Machine Learning for Diabetes and Mortality Risk Prediction From Electronic Health Records

Zakhriya N. Alhassan

A thesis presented for the degree of
Doctor of Philosophy at Durham University

Department of Computer Science
Durham University
United Kingdom

16th August 2021

Dedication

To the sake of Allah, all praise is due to him.

To my tender mother, Najebah Alsalman.

To my patient wife, Zainab Alhaddad.

To my lovely children, Fatima, Sakinah and Yahya.

To my brothers and sisters.

To my supportive supervisors, Dr. Noura Al Moubayed and Prof. David Budgen

Machine Learning for Diabetes and Mortality Risk Prediction From Electronic Health Records

Zakhriya N. Alhassan

Submitted for the degree of Doctor of Philosophy
August 2021

Data science can provide invaluable tools to better exploit healthcare data to improve patient outcomes and increase cost-effectiveness. Today, electronic health records (EHR) systems provide a fascinating array of data that data science applications can use to revolutionise the healthcare industry. Utilising EHR data to improve the early diagnosis of a variety of medical conditions/events is a rapidly developing area that, if successful, can help to improve healthcare services across the board. Specifically, as Type-2 Diabetes Mellitus (T2DM) represents one of the most serious threats to health across the globe, analysing the huge volumes of data provided by EHR systems to investigate approaches for early accurately predicting the onset of T2DM, and medical events such as in-hospital mortality, are two of the most important challenges data science currently faces. The present thesis addresses these challenges by examining the research gaps in the existing literature, pinpointing the un-investigated areas, and proposing a novel machine learning modelling given the difficulties inherent in EHR data.

To achieve these aims, the present thesis firstly introduces a unique and large EHR dataset collected from Saudi Arabia. Then we investigate the use of a state-of-the-art machine learning predictive models that exploits this dataset for diabetes diagnosis and the early identification of patients with pre-diabetes by predicting the blood levels of one of the main indicators of diabetes and pre-diabetes: elevated Glycated Haemoglobin (HbA1c) levels. A novel collaborative denoising autoencoder (Col-DAE) framework is adopted to predict the diabetes (high) HbA1c

levels. We also employ several machine learning approaches (random forest, logistic regression, support vector machine, and multilayer perceptron) for the identification of patients with pre-diabetes (elevated HbA1c levels). The models employed demonstrate that a patient's risk of diabetes/pre-diabetes can be reliably predicted from EHR records.

We then extend this work to include pioneering adoption of recent technologies to investigate the outcomes of the predictive models employed by using recent explainable methods. This work also investigates the effect of using longitudinal data and more of the features available in the EHR systems on the performance and features ranking of the employed machine learning models for predicting elevated HbA1c levels in non-diabetic patients. This work demonstrates that longitudinal data and available EHR features can improve the performance of the machine learning models and can affect the relative order of importance of the features.

Secondly, we develop a machine learning model for the early and accurate prediction all in-hospital mortality events for such patients utilising EHR data. This work investigates a novel application of the Stacked Denoising Autoencoder (SDA) to predict in-hospital patient mortality risk. In doing so, we demonstrate how our approach uniquely overcomes the issues associated with imbalanced datasets to which existing solutions are subject. The proposed model — using clinical patient data on a variety of health conditions and without intensive feature engineering — is demonstrated to achieve robust and promising results using EHR patient data recorded during the first 24 hours after admission.

Declaration

The work in this thesis is based on research carried out within the Innovative Computing Group (ICG) at the Department of Computer Science at Durham University, UK. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all the author's own work unless referenced to the contrary in the below.

Note on Publications Included in this Thesis

At the time of submission of this thesis, five main contributions of this thesis have been published or submitted for publication in journals or conferences.

Chapter 5: This chapter contains a published study in the ICANN conference.

- Alhassan Z, McGough AS, Alshammari R, Daghstani T, Budgen D, Al Moubayed N. *Type-2 Diabetes Mellitus diagnosis from time series clinical data using deep learning models.* In International Conference on Artificial Neural Networks 2018 Oct 4 (pp. 468-478). Springer, Cham. (Alhassan, McGough et al. 2018)

Chapter 6: This chapter contains a published study in the ICANN conference.

- Alhassan Z, Budgen D, Alessa A, Alshammari R, Daghstani T, Al Moubayed N. *Collaborative Denoising Autoencoder for High Glycated Haemoglobin Prediction.* In International Conference on Artificial Neural Networks 2019 Sep 17 (pp. 338-350). Springer, Cham. (Alhassan, Budgen, Alessa et al. 2019)

Chapter 7: This chapter contains two published studies. Both studies were published at the JMIR Medical Informatics journal.

- Alhassan Z, Budgen D, Alshammari R, Al Moubayed N. *Predicting Current Glycated Haemoglobin Levels in Adults From Electronic Health Records: Validation of Multiple Logistic Regression Algorithm.* JMIR medical informatics. 2020;8(7):e18963. (Alhassan, Budgen, Alshammari and Al Moubayed 2020)
- Alhassan Z, Watson M, Budgen D, Alshammari R, Alessa A, Al Moubayed N. *Improving Current Glycated Hemoglobin Prediction in Adults: Use of Machine Learning Algorithms With Electronic Health Records.* JMIR medical informatics. 2021;9(5):e25237.

Chapter 8: This chapter contains one published study in the ICMLA conference.

- Alhassan Z, Budgen D, Alshammari R, Daghstani T, McGough AS, Al Moubayed N. *Stacked denoising autoencoders for mortality risk prediction using imbalanced clinical data.* In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) 2018 Dec 17 (pp. 541-546). IEEE. (Alhassan, Budgen, Alshammari, Daghstani et al. 2018)

Description of these studies as presented in this thesis are largely as published or submitted. The references and notations have been altered, cross-references have been added and some stylistic changes have been made for the consistency throughout this thesis.

Chapter 5 The study included in this chapter was completed in partnership with McGough AS, Alshammari R, Daghstani T, Budgen D and Al Moubayed N. Alhassan Z was responsible

for preforming the majority of the work (implementing, validating and building machine learning models, analysing result and writing the manuscript). McGough AS, Budgen D, Al Moubayed N helped in designing and reviewing the manuscript. Alshammari R and Daghstani T helped in collecting and/or describing the dataset.

Chapter 6 The study included in this chapter was completed in partnership with McGough AS, Alshammari R, Daghstani T, Alessa A, Budgen D and Al Moubayed N. Alhassan Z was responsible for preforming the majority of the work (implementing, validating and building statistical and machine learning models, analysing result and writing the manuscript). McGough AS, Budgen D, Al Moubayed N helped in designing and reviewing/writing the manuscript. Alshammari R, Daghstani T, and Alessa A helped in collecting and/or describing the dataset.

Chapter 7 The studies included in this chapter were completed in partnership with Watson M, Budgen D, Alshammari R, Alessa A and Al Moubayed N. Alhassan Z was responsible for preforming the majority of the work (implementing, validating and building statistical and machine learning models, analysing results and writing the manuscript). Watson M helped in the explainability of the machine learning. Budgen D and Al Moubayed N helped in the designing and reviewing/writing the manuscript. Alshammari R, and Alessa A helped in collecting and/or describing the dataset.

Chapter 8 The study included in this chapter was completed in partnership with McGough AS, Budgen D, Alshammari R, Daghstani T and Al Moubayed N. Alhassan Z was responsible for preforming the majority of the work (implementing, validating and building statistical and machine learning models, analysing results and writing the manuscript). Budgen D and Al Moubayed N helped in designing and reviewing/writing the manuscript. Alshammari R, and Daghstani T helped in collecting and/or extracting the dataset.

Note on Publications Not Included in this Thesis

As well as the above studies, the following study has been published during the period of the research for this thesis.

- Alessa A, Faezipour M, Alhassan Z. *Text classification of flu-related tweets using fasttext with sentiment and keyword features.* In 2018 IEEE International Conference on Healthcare Informatics (ICHI) 2018 Jun 4 (pp. 366-367). IEEE. (Alessa, Faezipour and Alhassan 2018)

Acknowledgements

My deepest thanks go to my mother, wife, and lovely children. It means the world to have you by my side, come what may. Thank you for always being there. Special thanks go to my brothers (Ali, Sadeq, and Hassan) and sisters (Asma, Ashwaq, Azhar, Amnah, Ahlam, Abrar, and Sara) for always motivating and pushing me forward towards my goals.

My sincerest thanks go to my supervisors Dr Noura Al Moubayed and Prof David Budgen for their unlimited support in the preparation of this thesis. I thank you both for the invaluable advice, feedback, instructions, and scientific guidance you have provided throughout. I would also like to thank Dr A. S. McGough for his supervision in the first year of my PhD. Thanks also to all the members of Dr Noura's research group for all the helpful discussions and advice (especially, Tom Winterbottom and Matthew Watson).

I would like to acknowledge the contribution of King Abdullah International Research Center (KAIMRC) for providing me with access to the dataset under the approved projects "Diabetes Early Warning System, Research Protocol SP14/042" and "Finding the Common Related Diseases With Diabetes Using Data Mining Association Techniques, Research Protocol SP15/064" and extension project number RYD-17-417780-187503 to collect the most recent dataset.

I would also like to acknowledge the contributions of Prof Pali Hungin, Dr Majdy Aljassim, Naji Alhassan, Yousef Alhassan, and Dr Haidar Alhassan for their kind feedback about the clinical

aspects of the studies detailed in this thesis. My endless gratitude to Prof Riyad Alshammari, Dr Ali Al Muntasheri, Dr Ali Alessa, Ahmed Aldraihim, and Tahani Daghstani for their support in collecting the datasets used in this work. Thanks to Ahmed Alamri and Mohammed Alshihery for being true friends throughout.

Finally, I would like to thank the Saudi Cultural Bureau in Britain (SACB), Durham University (DU), and the University of Jeddah (UJ) for all their help in facilitating this project.

Chapter 6 Zakhriya Alhassan, David Budgen and Noura Al Moubayed are with Department of Computer Science, Durham University, Durham, UK. A. Stephen McGough is with Newcastle University, Newcastle upon Tyne, Tyne and Wear, UK. Ali Alessa is with Department of Information Technology Programs, Institute of Public Administration, Riyadh, Kingdom of Saudi Arabia. Riyad Alshammari and Tahani Daghstani are with College of Public Health and Health Informatics, Health Informatics Department, King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia and King Abdullah International Medical Research Center, Ministry of the National Guard - Health Affairs, Riyadh, Saudi Arabia.

Chapter 7 Matthew Watson is with Department of Computer Science, Durham University, Durham, UK.

Contents

List of Figures

List of Tables

List of Acronyms

ADA American Diabetes Association 21
AI Artificial Intelligence 24
ANFIS Artificial Neural Fuzzy Interference Systems 43
ANN Artificial Neural Networks 40
APACHE Acute Physiology, Age, Chronic Health Evaluation 22
AUR-ROC Area Under the Receiver Operating Characteristic 34
BiLSTM Bidirectional Long-Short Term Memory 30
BMI Body Mass Index 20
BPM Back Propagation Neural Network 42
bSMOTE Borderline Synthetic Minority Over-Sampling Technique 141
BUN Blood Urea Nitrogin 73
CHOL Total Cholesterol 73
DAE Denoising Autoencoder 32
EHR Electronic Health Records 2
ELM Extreme Learning Machine 42
EM Expectation Maximisation 39
FBS Fasting Blood Sugar 50
FFNN Feed Forward Neural Networks 40
GA Genetic Algorithm 39
GDA Generalised Discriminant Analysis 41

Chapter 1

Introduction

Prologue

This chapter proceeds by: (i) offering an introduction to this topic; (ii) assessing the benefits of using existing EHR data to predict health outcomes; (iii) explaining the motivation for investigating this area; (iv) assessing the contributions of the existing research; (v) outlining the scope of the current research project; and (vi) detailing the thesis' research outline and structure.

Specifically, both the introduction and the main research questions are presented in section 1.1. Details of the analytics (predictive models) driven by EHR are presented in section 1.2. Section 1.3 outlines the motivations for carrying out this research project. Sections 1.4 and 1.5 highlight the objective and main contributions of this thesis respectively. Section 1.6 defines the present study's scope. Finally, the research outline and structure are detailed in sections 1.7 and **??**.

1.1 Introduction

Nowadays, data science is revolutionising the healthcare sector by providing an invaluable set of tools that enable healthcare practitioners to understand and exploit complex Electronic Health Records (EHR) data (Hartskamp et al. 2019).

EHR systems are centralised, secure repositories of digitally stored patient data (ANSI-ISO 2005; Häyrinen, Saranto and Nykänen 2008). Those systems contain a range of time-stamped (longitudinal) data such as patient diagnoses, laboratory test results, vital sign readings, medication-prescription data, treatment plans, and physician notes. EHR data is recorded to support operational practices, not for research, and this does form a limitation upon its use. However, the large volumes of clinical data can be exploited (as a secondary use (Miotto et al. 2016)) to significantly enrich medical informatics to improve patient outcomes and provide cost-savings (Evans 2016).

With the aid of EHR systems, these datasets have come to represent an interesting frontier for knowledge extraction and hence improving decision-support systems. The past decade has witnessed a huge increase in the use of EHR data to enhance a variety of beneficial medical-decision-related support tasks (Coorevits et al. 2013), including the analysis of complex patterns of disease onset and the prediction of major medical events (Goldenberg, Nir and Salcudean 2019).

The task of predicting which patients are at risk of developing Type 2 Diabetes Mellitus (T2DM) — including the early identification of patients with pre-diabetes — is of great interest and forms a major challenge in the medical domain (Hasan et al. 2020; Wells, Lenoir et al. 2018). The early identification of such patients can help to mitigate the risk of patients developing diabetes-related complications and reduce the cost burden on health care resources. Nowadays, the Glycated Haemoglobin (HbA1c) blood test alone is suggested as a diagnostic test (World Health Organisation 2016) for patients at risk of pre-diabetes and diabetes. Additionally, elevated HbA1c levels can increase the risk of other serious health conditions such as cardiovascular disease in non-diabetic patients (Ackermann et al. 2011).

The prediction of patient in-hospital mortality risk is also important in the medical domain (Awad et al. 2017). Accurate measurement of patients' disease-acuity scores and the accurate prediction of mortality (or discharge type) events can improve healthcare services and patient survival outcomes (Alves et al. 2018). Therefore, to improve patient outcomes, physicians require tools that allow them to interpret clinical data outcomes quickly and accurately (Luo et al. 2016).

In recent years, machine learning tools have demonstrated an impressive capacity for the analysis and understanding of complex clinical data (Coorevits et al. 2013). Specifically, deep machine learning (Goodfellow et al. 2016) has been successfully adopted across a variety of medical applications including disease diagnosis, phenotyping, and medical-event prediction (H.-J. Jang and K.-O. Cho 2019; Durga, Nag and E. Daniel 2019).

EHR data provide an invaluable source of information that can be exploited to enhance predictive medical models. The present thesis aims to adopt state-of-the-art machine learning algorithms to overcome the challenges of utilising EHR datasets to address two major medical problems, (i) T2DM diagnosis using HbA1c levels (including the identification of pre-diabetic patients), and (ii) predicting patient in-hospital mortality risks.

1.1.1 Research Questions

This thesis investigates the use of recent machine learning technologies to address the two above-mentioned challenges by investigating the adoption of state-of-the-art machine learning approaches to exploit EHR to predict Glycated Haemoglobin (HbA1c) and patient mortality risk. To achieve this, we consider the following research questions:

- In what ways machine learning models can exploit EHR data to facilitate the early identification of patients at risk of diabetes via the prediction of Glycated Haemoglobin (HbA1c) levels? In particular:
  - Can machine learning models assist in predicting the levels of Glycated Haemoglobin (HbA1c) using typical EHR data for patients (diabetic and non-diabetic)?

  - How can machine learning assist with the early identification of patients with pre-diabetes via the prediction of elevated Glycated Haemoglobin (HbA1c) in patients with no history of hyperglycemia?
  - Can the use of temporal (time-series) data available in EHR help improve the elevated Glycated Haemoglobin (HbA1c) levels prediction using machine learning?
  - What is the impact of including temporal behaviour metrics on the importance of variables used to predict elevated Glycated Haemoglobin (HbA1c) levels?

- Can the risk of in-hospital patient mortality be measured accurately by machine learning models using EHR data?

1.2 Analytics Driven by Electronic Health Records Systems

Clinical studies that aim at providing analytics driven from EHR data differ from conventional studies, commonly employed with statistical models, that use traditional data collection approaches.

- **Data availability, collection effort and cost**

  In predictive machine learning-based medical studies, the required data is not normally available beforehand, leading to extra work during data collection. The effort and cost required to gather such data is dependent on the information to be collected and size of the intended sample population. In contrast, studies that utilise EHR data benefit from the availability of easily accessible pre-existing data; thus, minimal effort and cost are required to extract the required data from the selected EHR systems.

- **Prior clinical knowledge**

  In predictive medical studies, the involvement of clinical experts is crucial — especially in selecting, measuring, and recording the most appropriate variables. To elaborate further, it

is crucial that the variables chosen with respect to an evidence-based aetiology of the target condition. Thus, in studies based on conventional data collection approaches, prior clinical knowledge is required to undertake and record the readings from a patient blood sample at this stage. However, in the case of studies based on EHR data, variables for investigation can be selected with minimum clinical knowledge of how the data was measured as they are already available in EHR systems.

- **Research outcomes**

  Data collection procedures in conventional predictive studies are based on the intended research objectives and outcomes. In contrast, the studies driven by EHR data are not targeted at solving a specific research problem (this may or may not be useful). However, EHR data is currently used to solve a variety of problems and EHR datasets contain data from all patients visiting the hospital. Furthemore, such datasets tend to be more representative and better reflect real-world events compared to those used in conventional studies, as the latter depend on patient volunteers (Goldstein et al. 2017).

- **Data collection time**

  Conventional predictive studies usually collect longitudinal data about the target population. However, these studies require regular follow-ups of the target population. Therefore, sampling larger target populations requires more follow ups, which prolongs the data collection process. However, compared to conventional medical data collection approaches, using data from EHR systems involves only a fraction of the time required for data collection.

- **Dataset enrichness**

  While the data used in conventional predictive studies is limited to certain pre-selected variables, EHR datasets are usually enriched by hundreds of routinely collected and stored variables, which makes them an invaluable data resource for research.

1.3 Motivation

One of the aims of the use of decision-support-related tasks (i.e. predicting the risk of medical events or disease development) in the medical domain is to facilitate the early preventive interventions (Vogenberg 2009; Razzak, Imran and G. Xu 2020). This implies that such predictive tasks are invaluable in improving patient outcomes, supporting health service performance, minimising delays in diagnosing serious health conditions, and reducing health care expenditure.

Over the past few years, EHR data-driven medical analytics (and predictive tasks in particular) have become an interesting area of research that has led to pronounced improvements in the outcomes of a variety of medical applications. The wealth of EHR data provides an invaluable source of information to enrich medical predictive models.

The main motivation behind this work is using the wide availability of machine learning approaches and EHR datasets to solve complex problems and identify and extract latent data trends. Ultimately, this thesis is motivated by solving challenging classification problems related to diabetes and mortality risk prediction to improve the outcomes for patient and reduce the healthcare costs.

1.4 Research Objective

This research project aims to improve patient health care outcomes and reduce the burden to health care expenditure on treating T2DM via: the accurate and timely prediction of patients currently suffering from diabetes, the identification of pre-diabetic patients, and predicting the risk of in-hospital morality factors by utilising easily available EHR data by using state-of-the-art predictive methods (specifically, deep learning methods). The overall conceptual structure for this thesis is illustrated in Figure 1.1.

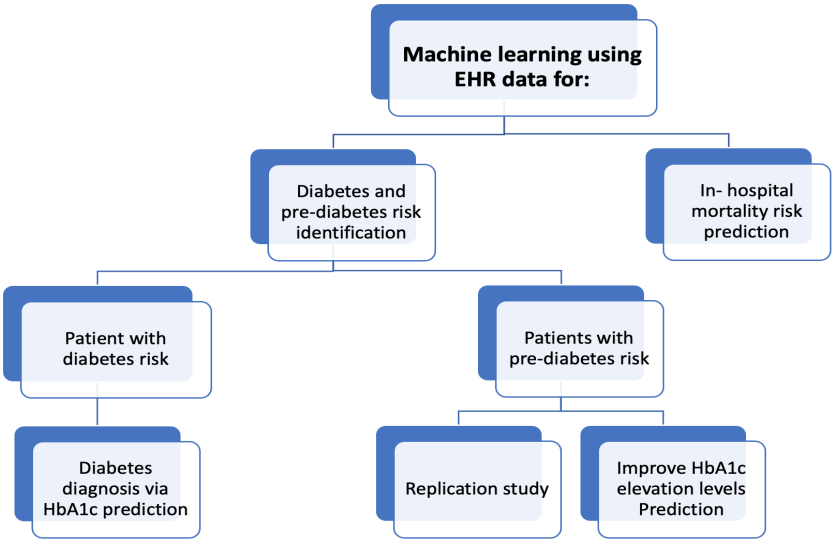


Figure 1.1: Overall conceptual structure for the thesis.

This thesis investigates using machine learning approaches with EHR data for diabetes and mortality risk prediction challenges due to their availability in KAIMRC dataset. However, those two challenges are not clinically aligned and hence have been investigated independently. Chapters 6 and 7 discuss diabetes risk prediction, while chapter 8 is aimed at investigating the in-hospital mortality risk prediction.

1.4.1 King Abdullah International Research Centre Dataset

The performance and accuracy of a particular machine learning model is heavily reliant on the training dataset employed. Therefore, to reach our research ambitions, a unique, large, and representative EHR dataset was selected: the King Abdullah International Research Centre (KAIMRC) dataset. The dataset was collected from Saudi Arabia where the diabetes is rapidly increasing (Abdulaziz Al Dawish et al. 2016). More details/statistics about the prevalence of diabetes in Saudi Arabia will be presented in section 5.3.1 of Chapter 5. This EHR dataset

includes time-stamp data that has been shown to maximise the predictive performance of the machine learning models we developed. More details about KAIMRC EHR dataset are provided in Chapter 5.

1.4.2 Diabetes Mellitus Diagnosis Using Glycated Haemoglobin

Diabetes is quickly becoming one of the worlds most problematic medical conditions. The most common form is Type 2 Diabetes Mellitus (T2DM) which accounts for 91% - 95% of all diabetes cases worldwide (International Diabetes Federation 2015). T2DM is difficult to diagnose because it does not demonstrate a clear set of immediately apparent clinical symptoms, often remaining undetected for long periods as a result of the slow onset of its identifiable symptoms (Beagley et al. 2014). Besides, the onset of T2DM can be completely symptomless, which further hampers its early diagnosis (International Diabetes Federation 2019). Glycated Haemoglobin (HbA1c) provides the basis for one of the most important blood tests used to assess the average glucose concentration in red blood cells (Peterson et al. 1998; Koenig et al. 1976). The International Expert Committee (IEC) and Diabetes Association (ADA) recommend effective clinical interventions for those patients with a HbA1c level of 6.5% or higher.

To achieve this, this thesis investigates an approach that utilises novel autoencoder-based deep learning model to identify the potential onset of T2DM by predicting abnormal levels of HbA1c (>=6.5%) in patient blood samples using the data normally stored in the EHR systems.

1.4.3 Pre-diabetes Diagnosis Using Glycated Haemoglobin Elevation Levels for Non-diabetic Patients

Patients with HbA1c levels lower than 6.5% are not completely excluded from a diabetes diagnosis as the range of elevation levels (5.7% >= HbA1c <6.5%) can indicate the likelihood of a future onset of diabetes (World Health Organisation 2011). Non-diabetic patients with an elevated HbA1c level (>=5.7%) are considered pre-diabetic and also have an increased risk of cardiovascular disease (Ackermann et al. 2011; Khaw et al. 2004).

Elevated blood HbA1c levels can cause chronic complications and lead to serious health conditions (Bonora and Tuomilehto 2011). Patients with HbA1c levels of 5.5% - 6.0% have up to 25% higher risk of developing diabetes, compared to patients with HbA1c levels of lower than 5.5% (X. Zhang et al. 2010). Furthermore, patients with HbA1c levels of 6.0% or higher are 50% more likely to develop T2DM over the next 5 years. By comparison, patients with HbA1c levels lower than 5.0% are 20 times less likely to develop T2DM.

The HbA1c test helps physicians decide how frequently patients need to undertake clinical screening for T2DM (Edelman et al. 2004). It is hoped that providing an accurate method of identifying patients at risk of pre-diabetes via the prediction of elevated HbA1c levels will ultimately assist physicians to prevent the development of diabetes and mitigate the long-term complications of this disease.

On this basis, predicting elevated levels of HbA1c in patients with no history of hyperglycemia will be investigated in this thesis by: (i) replicating a recent study using statistical models on the KAIMRC dataset; (ii) improving prediction performance by adopting advanced learning models that incorporate larger dimensional time-series data available in the KAIMRC dataset to identify pre-diabetic patients by forecasting their likelihood of having elevated HbA1c levels (>= 5.7).

1.4.4 In-hospital Mortality Risk Prediction

The accurate prediction of mortality risk can help health care services improve patients' likelihood of survival (Alves et al. 2018; Awad et al. 2017). Therefore, physicians require timely and effective tools to help them interpret clinical data and improve patient outcomes (Luo et al. 2016). Thus, the early prediction of in-hospital mortality risk is an area of interest to us in the current research project.

To achieve this research aim, we use an autoencoder-based approach to mortality risk prediction. In brief, patient mortality prediction is formulated into a binary classification problem for predicting an in-hospital patient's mortality risk in general (regardless of the health condition types) and after only 24 hours of patient admission using KAIMRC dataset.

1.5 Contributions

This thesis provides several contributions. The main contributions are as follows.

- Introducing a unique, large, representative and time-stamped medical dataset extracted from the KAIMRC EHR systems containing data for the period from 2010 - late 2018.
- Adopting state-of-the-art predictive machine learning algorithms featuring novel modelling structures to meet the challenges of accurate classification of diabetes and mortality risk from EHR data. This includes adopting:
  - Collaborative autoencoder along with Multilayer perceptron models to accurately predict the HbA1c diabetes/pre-diabetes levels
  - Stacked denoising autoencoders for in-hospital mortality risk prediction.
- Replicating and evaluating recent studies that predict elevated HbA1c levels and adopting state-of-the-art machine learning approaches with the KAIMRC dataset using machine learning. This includes:
  - Performing a differentiated replication study to validate, evaluate, and identify the strengths and weaknesses of the predictive models when used to forecast the levels of HbA1c using the KAIMRC dataset.
  - Investigating employing advanced machine learning approaches for elevated HbA1c levels prediction.
- Investigate how time-series data from the KAIMRC EHR dataset affects the performance and the importance of different features using the employed diabetes-risk-related predictive models. This includes:
  - Utilising the patient's EHR longitudinal data to improve the performance of the predictive models used.

  - Adopting state-of-the-art explainable machine learning techniques to explain the classification decisions made by the machine learning models employed and rank the importance of the features used.

- Presenting a unique interpretative approach to overcome the challenges of using imbalanced datasets for mortality risk prediction.

1.6 Scope

Two major factors have shaped the scope of this research; (i) the limitations of the KAIMRC dataset, and (ii) the lack of available datasets and similar studies that employed machine learning models to process EHR data for predictive medical analytics in general, and for diabetes and mortality risk prediction specifically.

1.6.1 Limitations of the KAIMRC dataset

- **Dataset Extraction Phases**

  The dataset upon which the present study is based was originally extracted in two stages for two separate research projects (as discussed in chapter 5). Therefore, differences exist between the two parts of the dataset collected in each phase. The main differences are:

  - The omission of several important variables in both parts of the dataset. There were some variables available in one of the parts of the dataset but not in the extended part.
  - The first part of the dataset was extracted from EHR data across three hospitals while in the second, the data were collected from only two of the three hospitals used in the first phase.

  These issues (will be discussed in details in Chapter 5) limited the investigated classification models and the investigations of the outcomes of this research, to be applied on, and discussed with respect to the complete dataset (two parts).

- **Type of the Collected Dataset**

  The dataset obtained from the KAIMRC was originally collected for diabetes-related projects (refer to section 5.3.2 of Chapter 5). Therefore, to exploit this, diabetes prediction through HbA1c blood test using data from EHR systems was adopted for the main focus of this thesis (the HbA1c values were available in both parts of the obtained dataset). However, due to the availability of the discharge type (in-hospital mortality) details for patient visits in part 1 of the KAIMRC dataset, this thesis also investigated the scope for predicting the in-hospital mortality as a second direction for investigation using EHR data (the discharge type details were not available in part 2).

  EHR systems contain both structured data (vital sign values and lab test results) and unstructured data (physician's notes). However, only structured data is extracted from the EHR systems for this research, which contains basic patient data (age and gender), lab test results, diagnoses, vital signs readings, and prescribed medications. Therefore, because only structured data is available, and due to the nature of the clinical datasets, this limits the extent to which the project can investigate the use of machine learning techniques.

- **Data Compliance Requirements**

  Each organisation has its own data compliance regulations; data sensitivity is one of the main affecting data compliance in the medical domain. Therefore, to meet the dataset provider's compliance requirements for this research, only specific and limited data can be harvested from the dataset. Examples of such restricted patient data include patients' smoking status and family information; these data were not included when collecting the KAIMRC dataset.

- **Data Population**

  Only data from adult patients was extracted from the KAIMRC dataset; patients under 18 were excluded. Also, it is important to highlight that the hospitals from which the data were collected are private (i.e. not public) facilities where patients are limited to employees of the Ministry of National Guard and their families. However, employees of

some companies in Saudi Arabia (Saudi Aramco and Sabic) have recently been permitted to use the National Guard hospitals.

1.6.2 Lack of Suitable Public Datasets

At the time of writing, no suitable medical datasets were publicly available for use in the project; this is due to the previously mentioned privacy and confidentiality restrictions that usually apply to such datasets. These restrictions limited the predictive models to be utilised and generalised using different population.

1.6.3 Lack of Similar Studies on this Topic

Although machine learning models have shown an impressive capacity for analysing and understanding complex data across a wide variety of applications, there is still much to do in terms of applying machine learning using EHR to benefit patients in the medical domain (Harerimana et al. 2019). Although the use of EHR data is an increasingly popular research area nowadays, few studies exploit EHR data in terms of utilising HbA1c levels to predict the risk of diabetes (as discussed in the related work; Chapter 4). Thus, comparing the performance achieved by the models employed in this work with those developed by others is limited.

1.7 Research Outline and Structure

This thesis is organised as follows:

Chapter 2 This chapter provides a discussion of the healthcare context in relation to the aims of the present study. This chapter begins by briefly introducing T2DM, followed by an overview of HbA1c; it concludes by briefly describing the concept of in-hospital mortality risk.

Chapter 3 After introducing the main clinical concepts related to our research aims, this chapter provides a description of the predictive models used in this thesis. It provides a brief introduction

to the machine learning approaches employed throughout this thesis and the measures used to evaluate the performance of the models employed.

Chapter 4 After presenting the relevant healthcare-context-related background knowledge and the employed predictive models, this chapter lists and discusses the main existing predictive models and the datasets used in T2DM diagnosis, elevated HbA1c levels, and in-hospital mortality risk predictions. Most importantly, it concludes by discussing the gaps identified in the existing literature.

Chapter 5 Datasets can play a major role in statistical and machine learning-based analytics. This chapter introduces the King Abdullah International Research Centre (KAIMRC) dataset and outlines the collection and preparation of this dataset. It also discusses the characteristics of this dataset and the issues that affect it. Finally, it introduces the sampling approaches used to create the subsets obtained from the KAIMRC dataset.

Chapter 6 T2DM is the most common type of diabetes; due to its unclear clinical symptoms, it can remain undetected until it becomes a serious health issue. Thus, facilitating the early identification of T2DM patients can help provide better patient outcomes and reduce cost burden to healthcare services. This chapter investigates the application of novel unique deep learning models utilising HbA1c levels to predict the risk of diabetes.

Chapter 7 This chapter investigates the identification of patients with pre-diabetes using machine learning. It begins by performing a differentiated replication study to identify the strengths and weaknesses of the predictive models to forecast HbA1c levels across different populations. Next, it investigates how HbA1c prediction can be improved by employing advanced machine learning approaches, incorporating longitudinal data, and the inclusion of more features.

Chapter 8 Predicting patient mortality risk is a major concern for physicians. To address this, this chapter employs a unique approach to investigate the prediction of in-hospital mortality risk using imbalanced clinical data extracted from EHR systems. This chapter also details the methodology used to tackle the problem of data imbalance.

Chapter 9 This chapter discusses the merits and limitations of this study's contributions to the literature. It also outlines potential future research avenues for using EHR data in predictive machine learning models.

Chapter 10 This chapter provides a review, summary, and conclusion of the key findings of this thesis.

Finally, it is important to highlight that each of the core chapters (6, 7 and 8) that address the research questions of this thesis represents an independent study with regards to the objectives, datasets and methods applied.

Chapter 2

Healthcare Context

Prologue

In this chapter, we outline the background on the main clinical concepts relevant to this thesis. An overview of diabetes is presented, focused on Type 2 Diabetes Mellitus (T2DM). The global impact of diabetes is briefly shown. The second section provides an introduction of Glycated Haemoglobin (HbA1c) and the health complications it may cause. The last section briefly details in-hospital mortality risk and the various scoring systems used.

2.1 Diabetes Mellitus

Diabetes Mellitus (DM) is a serious and chronic global health problem (Y. Wu et al. 2014). DM occurs when the body fails to produce enough insulin or when the cells of the body fail to respond to the insulin produced (International Diabetes Federation 2017). Insulin is a hormone that is produced by the pancreas and it is responsible for transporting glucose from the blood into the body's cells to be used as a source of energy (International Diabetes Federation 2015). A raised level of blood glucose leads to damages in many parts of the body's systems (i.e. heart

and kidneys), causing serious health complications (World Health Organisation 2016). Diabetes falls into three main types (International Diabetes Federation 2017):

- Type 1 Diabetes Mellitus (T1DM): occurs when the immune system of the body attacks the cells that produce insulin in the pancreas (beta cells), resulting in insufficient amounts of, or even no insulin being produced.
- Type 2 Diabetes Mellitus (T2DM): occurs when the body fails to respond to the insulin produced. As a result, blood glucose levels increase, which can lead to serious complications such as cardiovascular, diabetes-related kidney disease, eye issues and other diseases.
- Gestational Diabetes: occurs due to hormonal changes during pregnancy that increase the body's resistance to insulin.

Type-2 Diabetes Mellitus (T2DM) is a common and widely growing global medical condition and is recognised as one of the world's fastest-growing chronic medical condition (International Diabetes Federation 2019). T2DM accounts for 91% to 95% of all diabetes cases worldwide (International Diabetes Federation 2015). The estimated number of diabetic patients worldwide was 415 million in 2015, with 46% undiagnosed. The number of patients with diabetes increased to 463 million in 2019 (International Diabetes Federation 2019). In the UK, current estimates show that at least 1 in 16 people have diabetes and four million people are affected in total. Furthermore, it is expected that 700 million people will have diabetes by 2045. Table 2.1 shows the distribution of diabetes estimated by gender. It is thought that this steady increase is caused by: population growth, increasing average physical inactivity and obesity and being overweight (World Health Organisation 2014).

The number of people with borderline diabetics (pre-diabetic) is also increasing extremely rapidly. Estimates indicate that 374 million people are currently at a high risk of developing T2DM (R. Williams et al. 2020). For instance, recent estimates show that 35.3% of UK adults are pre-diabetic (Wise 2014). This means that 1 in 3 UK adults are borderline diabetic (pre-diabetic).

Table 2.1: Estimated global number of adult patients with diabetes.

	2019		2030		2045	
Gender	Number of diabetic patients (millions)	Prevalence (%)	Number of diabetic patients (millions)	Prevalence (%)	Number of diabetic patients (millions)	Prevalence (%)
Male	240.1	9.6	296.7	10.4	357.7	11.1
Female	222.9	9.0	281.8	10.0	342.5	10.8

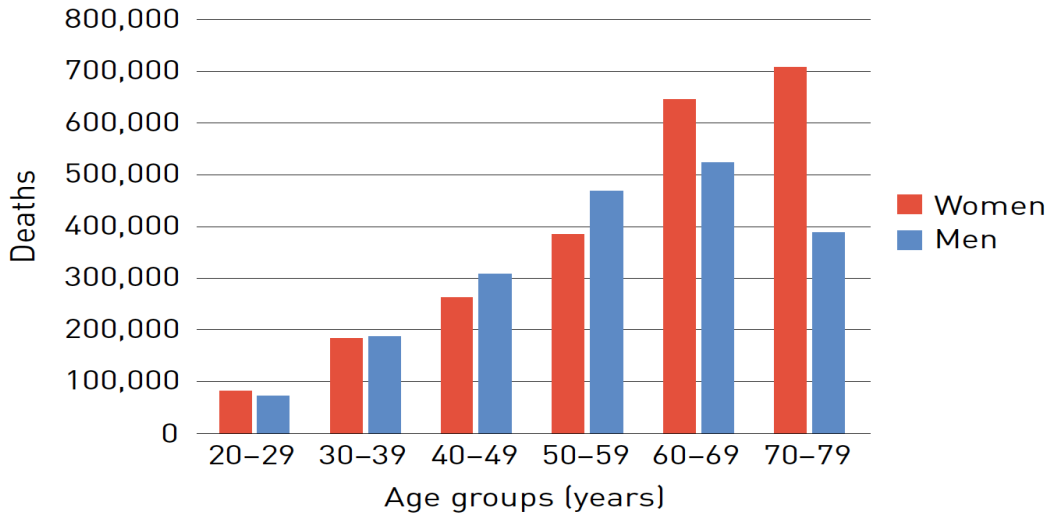
(The estimations shown in this table are adapted from the IDF Atlas report for 2019 (International Diabetes Federation 2019)).

In the United States, more than 60 million are pre-diabetic (which is also one third of the adult population). According to the Centers for Disease Control and Prevention (CDC), only 7% of those patients were told that they are pre-diabetic. The studies show that 40% of the those patients will develop diabetes within 10 years (Ackermann et al. 2011).

Unfortunately, patients suffering from diabetes are highly likely to develop serious and complicated health complications related kidney, vision, cardiovascular, and others. Diabetes is the main cause of renal failure, amputation and loss of vision (Salman, AlSayyad and Ludwig 2019). By the end of 2019, 4.2 million deaths were caused by diabetes worldwide. Figure 2.1 shows the distribution of diabetes related deaths over adult patients by age and gender.

Besides the human cost of diabetes, treating diabetic patients accounts for substantial portion of health providers' resources. For instance, the world health care expenditure on the treatment for patients with diabetes alone costs USD $760 billion per year, which represented 10% of the total world health expenditure in 2019 (R. Williams et al. 2020). Thus, the early identification of patients with diabetes can help physicians to plan preventive interventions that can delay or prevent diabetes complications. As a related benefit, this could arguably reduce the huge health care expenditure currently used on treating diabetes.

Figure 2.1: Number of diabetes-related deaths in adult patients by age and gender. Adapted from the IDF Atlas report for 2019 (International Diabetes Federation 2019).



2.2 Glycated Haemoglobin

Haemoglobin, when combined with blood glucose, forms Glycated Haemoglobin, referred to as HbA1c (Peterson et al. 1998; Koenig et al. 1976). Glycated Haemoglobin (HbA1c) provides the basis for one of the most important blood tests used to indicate the average glucose concentration in red blood cells (Peterson et al. 1998; Koenig et al. 1976).

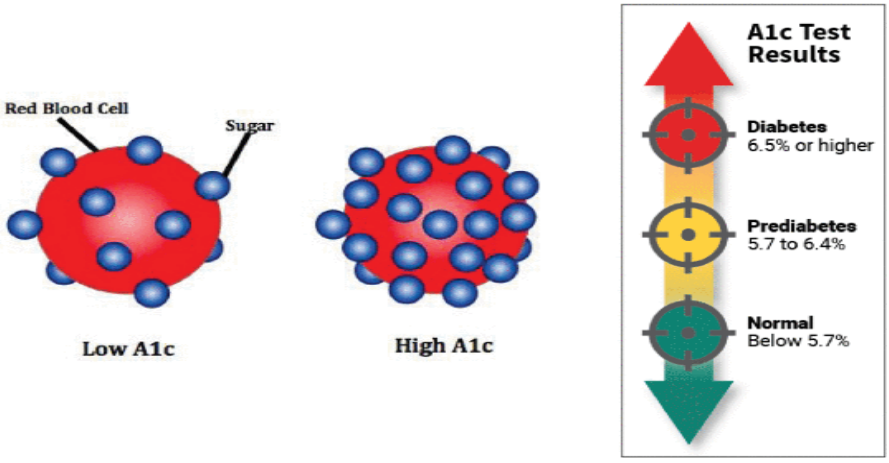
The level of HbA1c is strongly related to the average glucose concentration in the blood and the life span of the red blood cells. Under normal conditions, red blood cells of the human body can last for two to three months before being reproduced. Therefore, the level of HbA1c can indicate the average levels of blood glucose over the whole period of the life span of the red blood cells (Larsen, Hørder and Mogensen 1990; A. D. Pradhan et al. 2007). This can provide physicians with an important long-term measure for blood glucose levels (Ackermann et al. 2011).

One of the methods of diagnosing diabetes is to use the fasting blood sugar (FBS) test (American

Diabetes Association 2014). The FBS test can provide a measure of short-term blood glucose level but requires the patient to undertake overnight fasting prior to the test. However, the HbA1c test can also provide physicians with reliable readings to enable a patient's hyperglycemia to be monitored and managed. The HbA1c blood test can provide an overall average measurement of blood glucose levels over the preceding two to three months (long-term) and can be taken without requiring the patient to undertake overnight fasting prior to the test. HbA1c levels are least affected by short term illnesses that are correlated with plasma glucose levels. Thus, the HbA1c test is an attractive option for both patients and practitioners for measuring glucose levels in the blood.

The cut-off point for a diagnosis of diabetes when using the HbA1c blood test is a concentration of 6.5% and higher (World Health Organisation 2011). However, patients with less than 6.5% are not completely excluded from a diabetes diagnosis as the range of elevated levels (5.7>= HbA1c <6.5), can act as cut-off point for a diagnosis of pre-diabetes and the future onset of diabetes. Figure 2.2 shows the Glycated Haemoglobin(HbA1c) test results in a diabetes diagnosis.

Figure 2.2: Glycated Haemoglobin(HbA1c) in diabetes diagnosis. Adapted from (Islam, Qaraqe and Belhaouari 2020).



Recent studies have shown that the HbA1c test can be used as an indicator for diagnosing T2DM, especially when combined with vital signs such as Body Mass Index (BMI) (Gerstein et al. 2007;

J. Kälsch et al. 2015). Ackermann et al. (2011) suggested using the HbA1c test as a measure for identifying adults at a greater risk of developing T2DM in the future. Therefore, HbA1c can act as an early predictor for the potential development of T2DM (A. D. Pradhan et al. 2007).

The International Expert Committee (IEC), along with with members of the American Diabetes Association (ADA), the European Association for the Study of Diabetes, and the International Diabetes Federation (IDF) (International Expert Committee 2009; American Diabetes Association 2014), recommends using the HbA1c test to evaluate the adults with a high risk of diabetes (Ackermann et al. 2011). HbA1c levels are also related to chronic complications (Bonora and Tuomilehto 2011). The IEC recommends effective clinical intervention for patients with a HbA1c level of 6.5% or more.

Non-diabetics people with an elevated HbA1c level, higher than 5.7%, are also at an increased risk of cardiovascular disease (Khaw et al. 2004; Ackermann et al. 2011). The HbA1c test helps to predict which patients are likely to develop T2DM in the future. It also helps physicians decide how frequently patients need to undertake clinical screening for T2DM (Edelman et al. 2004). The earlier the diagnosis of T2DM, the better the chance of delaying (and possibly preventing) long-term complications (International Diabetes Federation 2017).

Research has shown that reducing HbA1c levels can reduce the possibility of developing serious complications in diabetic patients (Stratton et al. 2000; Group 1998). Lowering the HbA1c level by 1% for diabetic patients can help reduce the risk of developing heart failure by 16%, cataracts by 19%, retinopathy and kidney disease by 25%, and death caused by vascular diseases by 43%. Hence, close monitoring of HbA1c levels is recommended for diabetic patients and also for those with the potential for developing diabetes (Khaw et al. 2004).

2.3 In-hospital Mortality Risk

The definition of in-hospital mortality is the occurrence of a patient's death while receiving treatment at hospital (Awad et al. 2017). The in-hospital mortality rate can be used as an

indicator of the quality of the healthcare services provided to the patients (Goodacre, Campbell and Carter 2015). The number of in-hospital deaths in the United States in 2010 was 715,000 cases, which represents 2.03% of the total number of hospitalised patients (35.1 million), which was a decrease of 8% compared with those in 2000 (M. J. Hall, Levant and DeFrances 2013). Reducing in-hospital mortality rates is one of the main goals of hospitals and desired by patients (J. Wright et al. 2006). Therefore, several scoring systems have been developed to measure in-hospital patient mortality risk. These scoring systems are used by physicians to identify patients early who have a high risk of mortality to allow for quick and timely interventions.

The calculation of in-hospital mortality risk score for patients is usually evaluated by physicians using traditional scoring systems, such as Simplified Acute Physiology Score Mellitus (SAPS), Euro- SCORE, Acute Physiology, Age, Chronic Health Evaluation (APACHE), Mortality Probability Models (MPMs), Sequential Organ Failure Assessment (SOFA) score and Pediatric Risk of Mortality (PRISM) (Doig et al. 1993; Knaus et al. 1991; J.-L. Vincent, De Mendonça et al. 1998). Physicians use these systems to assess a variety of different clinical conditions, which differ in the clinical and biological variables used but all aim to provide an accurate calculation of the in-hospital mortality score.

One of the most common traditional scoring system physicians use as an indicator for patient acuity status is the Simplified Acute Physiology score (SAPS) (Keegan, Gajic and Afessa 2012). To date, three versions of SAPS exist: SAPS I, SAPS II, and SAPS III. The early version of SAPS is calculated manually after the first 24 hours of patient admission using 14 clinical and biological variables (J. G. Le et al. 1984). The SAPS system uses minimum and maximum variable values. Different imputation approaches are used in case of missing data such as replacing a maximum value with a minimum value in the case of a missing maximum value and vice versa.

The Mortality Probability Model (MPM) introduced by Lemeshow et al. in 1985 (Awad et al. 2017) uses seven variables to be collected at the time of admission and then 24 hours afterwards. The value of these variables is then converted into a categorical value (affirmative or negative). Another version of MPM II was introduced in 1993 (Lemeshow et al. 1993). MPM II measures the

in-hospital mortality score using two models at zero hour (MPM_0) and 24 hours (MPM_{24}) after a patient's admission. MPM_0 uses 15 variables to be collected at the time of admission while MPM_{24} requires five variables from MPM_0 in addition to eight more variables to be collected 24 hours after admission. Next, EuroSCORE is used to calculate a patient's mortality risk after heart surgery (Nashef et al. 1999). The SOFA score is designed to calculating a mortality score for patients admitted to hospital with one or more of six specific health conditions (Awad et al. 2017).

Several studies have compared the performance of the above models (Gilani, Razavi and Azad 2014; Kramer, Higgins and Zimmerman 2014). However, J.-L. Vincent and Singer (2010) highlight that these scoring systems should be combined as complementary systems. The currently in-use mortality scoring systems differ in many ways; for example, type and number of variables that need to be collected, collection time, imputation approaches, and calculation methods. However, the majority of these systems are not used for the early prediction of mortality. In practice, the use of traditional systems does not provide strong support for accurately calculating in-hospital mortality risk (Awad et al. 2017).

An automated approach that can systematically measure in-hospital mortality using patient observations stored via EHR using advanced medical predictive models can help clinicians provide patients with timely interventions. Thus, these predictive models are aimed at improving health care services to better support the patient survival.

Epilogue

In this chapter, we have introduced the main healthcare concepts that are related to this thesis. Type 2 Diabetes Mellitus (T2DM) was briefly described followed by an overview about Glycated Haemoglobin (HbA1c) and the in-hospital mortality scoring systems. The next chapter will provide the details about the predictive machine learning approaches employed to address the objectives of this thesis.

Chapter 3

Predictive Machine Learning Approaches

Prologue

Machine learning is a field of Artificial Intelligence (AI) that involves automatically learning and extracting of knowledge from raw samples of data without explicit programming in order to solve specific problems (Goodfellow et al. 2016; Alpaydin 2020). With the help of Electronic Health Records (EHR), studies using clinical data are at the forefront of machine learning research. In recent years, machine learning models have demonstrated a remarkable capacity for understanding and analysis of complex clinical data in a variety of medical applications (Coorevits et al. 2013), such as identification of patients with diabetes and pre-diabetes as well as in-hospital mortality risk measuring.

Since the KAIMRC dataset (detailed in Chapter 5) contains structured (time-series) data, this work focused on investigating the application of several deep learning approaches that can be used with structured and/or temporal (time-series) data. The MLP, RNN (this includes LSTM,

BiLSTM, and GRU), and autoencoder approaches are used in this thesis. Specifically, the MLP is used in Chapters 6, 7, and 8. The RNN approaches are used in Chapters 7 and 8. The autoencoders are used Chapters 6 and 8.

This chapter provides a brief overview of the predictive machine learning approaches employed to address the objectives of this thesis. This chapter proceeds by first discussing predictive technologies that are employed in this thesis; namely, deep machine learning technologies followed by the conventional machine learning approaches.

3.1 Deep Learning Approaches

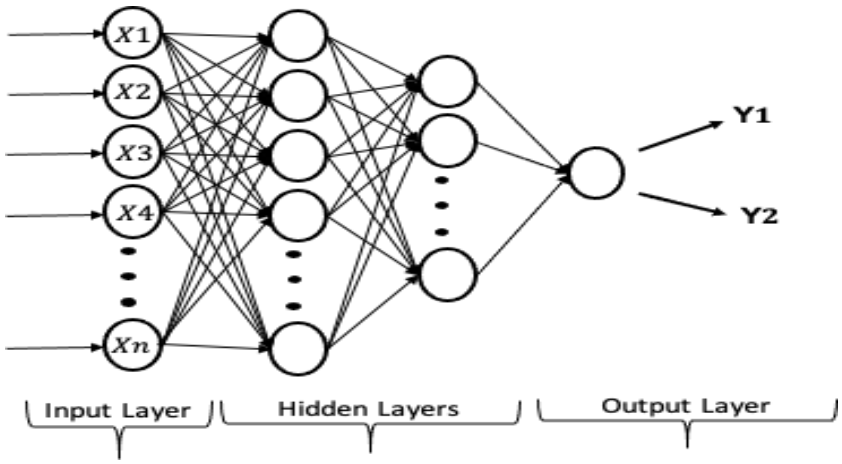
Deep learning is a type of machine learning inspired by the neural system of the human brain. Deng and D. Yu (2014) define deep learning as "a class of machine learning algorithms that uses multiple layers of nonlinear processing units for feature extraction and transformation". In general, machine learning algorithms can be categorised as supervised and unsupervised learning algorithms. Supervised algorithms learn the mapping between input and the associated output/target feature(s) of a dataset (labelled data), while unsupervised algorithms create internal representations from the structure and the distributions of data samples that are not associated with output/target feature(s) to identify the properties/classifications of these samples (unlabelled data) (Goodfellow et al. 2016). The process of feeding the machine learning model with data to learn from is referred as "training" while qualifying/assessing the performance of the trained/built model using unseen samples (not part of the training set) is referred as "testing" (Kuhn, K. Johnson et al. 2013).

Recently, deep learning models have been shown successful and useful in many areas such as natural language processing, image segmentation, object detection, bioinformatics, and gaming (Goodfellow et al. 2016). Deep machine learning modelling has become feasible for two main reasons: (i) the availability of large datasets that can help train them; (ii) the availability of powerful yet inexpensive computational technologies (J. Schmidt et al. 2019; Holder 2018).

3.1.1 Multi-Layer Perceptron

Multi-Layer Perceptron (MLP), also known as a feed-forward neural network, is one of the most common deep learning approaches. Making use of artificial neural networks, it is mainly used to address supervised learning problems (learning the mapping between input and output variable(s) to predict unseen data (Cunningham, Cord and Delany 2008)). The general structure of most MLP models contains an input layer, hidden layer or layers, and output layer. Figure 3.1 provides an example of an MLP structure for a binary classification problem. MLP models learn the dependencies between the input layer (the features or variables) and the output layer (the classification layer) using a fully connected hidden layer in-between.

Figure 3.1: General structure of Multi-layer Perceptron (MLP).



An example of multi-layer perceptron (MLP) model consisting of an input layer, two hidden layers and an output layer.

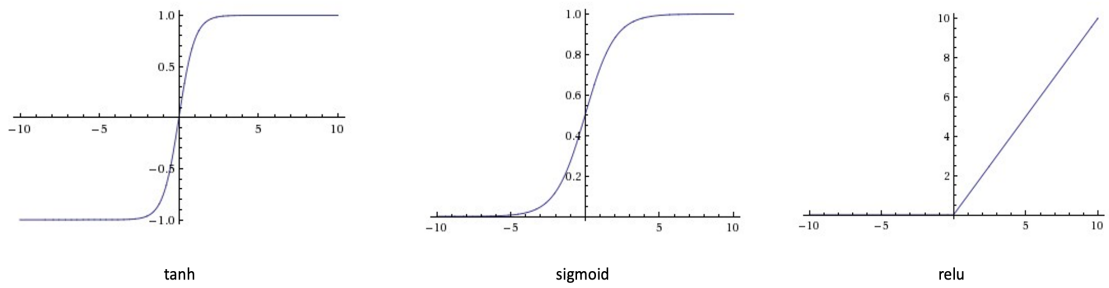
The layers (including the hidden ones) contain neurons that are connected to the neurons of the next and previous layers via weights and functions. MLP uses a backpropagation algorithm to update the weights and biases within the hidden layers to minimise the output error (or loss) (Gardner and Dorling 1998; LeCun, Bengio and Hinton 2015). The weights automatically select important information from the inputs. The neurons are represented by real numbers (approximated by floats in computers) that are multiplied by the weights, which are then summed

with a bias value (Eq. 3.1).

$$Y = \sum(x * weight) + bias \tag{3.1}$$

Then, the summed value is transformed into an output using activation functions and sent to the following layer. The activation functions are used to add a non-linear property to the output of a neuron of a layer. Adding such a non-linear property helps the neural network solve complex problems. Popular activation functions include sigmoid, tanh and relu (Szandała 2021) (shown in Figure 3.2).

Figure 3.2: Sigmoid, tanh and relu activation functions. Adapted from (Karpathy et al. 2016).



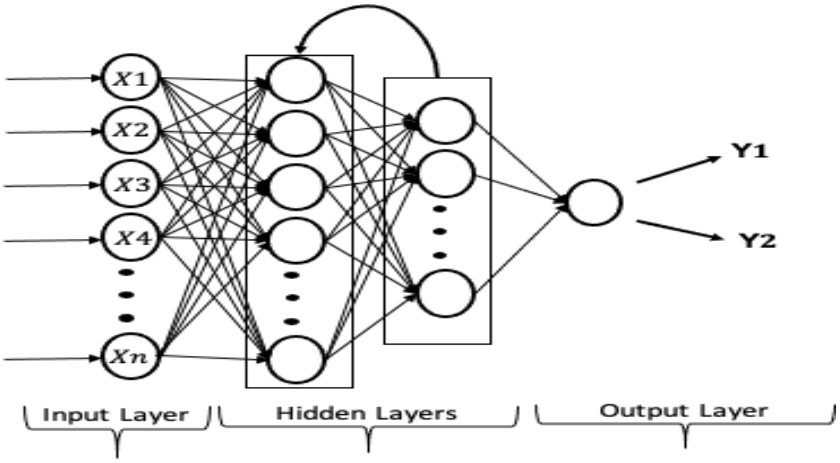
The loss function (also called cost function) is used to measures the deviation between the predicted output of the model and the ground truth (Bengio, Goodfellow and Courville 2017). According to the output of the this function, the weights of the neural network are changed via optimisation methods. These optimisation methods are used to tune the weights and the learning rate of the neural network to minimise the losses (Goodfellow et al. 2016).

3.1.2 Recurrent Neural Network

Recurrent Neural Network (RNN) were derived from MLPs or feedforward neural networks (LeCun, Bengio and Hinton 2015). RNNs (and its variants) have achieved unprecedented

accuracy in many domains with sequential data such as text for natural language processing. Unlike other deep learning methods, RNNs make use of memory cells allowing the previous output to influence the state for the next output. Figure 3.3 shows an example of the structure of an RNN for a binary classification problem.

Figure 3.3: The Structure of recurrent neural network (RNN).



*This figure shows an example of recurrent neural networks (RNNs) model consisting of an input layer, two hidden layers and an output layer.

A hidden state in an RNN network is calculated based on the an input x at time t and the hidden state of previous input at time $t-1$ (Eq. 3.2):

$$h_t = f(h_{(t-1)}, x_t) \tag{3.2}$$

Long-Short Term Memory

Long-Short Term Memory (LSTM) is an extended version of a recurrent neural network (RNN) that has achieved unprecedented accuracy in many domains with time series data that require long and short dependencies (LeCun, Bengio and Hinton 2015). Similar to the RNN, the LSTMs make use of memory cells to influence the state for the next output. LSTMs can be trained using an input formed in future time direction (Hochreiter and Schmidhuber 1997).

In practice, RNNs have demonstrated limited performance when learning from sequences with long-term dependencies (Bengio, Simard and Frasconi 1994). This is mainly caused by limitations in the gradient decent, as the gradient tends to either vanish or explode when modelling long dependencies. Hochreiter and Schmidhuber (1997) addressed this problem by introducing the LSTMs. LSTMs, use a sophisticated structure with multiple cell and gated units (i.e. forget and input gates) to cope with learning from long-term dependencies. This is as follows:

$$f_t = \sigma(W_f.[h_{(t-1)}, x_t] + b_f) \tag{3.3}$$

$$i_t = \sigma(W_i.[h_{(t-1)}, x_t] + b_i) \tag{3.4}$$

$$\tilde{C}_t = tanh(W_C.[h_{(t-1)}, x_t] + b_C) \tag{3.5}$$

$$C_t = f_t \times C_{(t-1)} + i_t \times \tilde{C}_t \tag{3.6}$$

$$o_t = \sigma(W_o[h_{(t-1)}, x_t] + b_o) \tag{3.7}$$

$$h_t = o_t \times tanh(C_t), \tag{3.8}$$

where f represents the forget gate of the cell with a sigmoid activation function σ, the weight W, and the learned bias b (Eq. 3.3). i is the input gate (Eq. 3.4) which is used in combination with a non-linear (tanh) layer $\tilde{C}$. $\tilde{C}$ is the new value for cell state (Eq. 3.5). The update state value C becomes the sum of the products of the old state $C_{(t-1)}$ by f_t, learns which features to ignore or preserve through time, and the new value $\tilde{C}$ multiplied by the input gate value i_t (Eq. 3.6). Finally, o is the output of the sigmoid gate which is used with the cell state C to produce the final decision (Eq. 3.7 and Eq. 3.8).

Bidirectional Long-Short Term Memory (BiLSTM) is an extension of the LSTM introduced in 1997 (Schuster and Paliwal 1997). BiLSTM can be trained by adding a duplicate layer to the input layer using past time direction (reversed input).

Gated-Recurrent Unit (GRU)

Similar to LSTMs, a Gated Recurrent Units (GRU) is used to deal with long-term dependencies. The main difference is that GRUs merge the forget and input gates into one unit gate called the update gate (Chung et al. 2014; K. Cho et al. 2014). This means that previous memory is retained based on the size of the new dependencies (input). GRUs do not have a protected hidden cell state, which allows full access to the corresponding allocated memory content. A GRU is defined as follows (Chung et al. 2014):

$$z_t = \sigma(W_f \times [h_{(t-1)}, x_t]) \tag{3.9}$$

$$r_t = \sigma(W_r.[h_{(t-1)}, x_t]) \tag{3.10}$$

$$\tilde{h}_t = tanh(W_C.[r_t \times h_{(t-1)}, x_t]) \tag{3.11}$$

$$h_t = (1 - z_t) \times h_{(t-1)} + z_t \times \tilde{h}_t, \tag{3.12}$$

where z and r represent the update gate and the reset gate values respectively. These gates are calculated in a similar way to calculating the input gate and the forget gate of LSTMs. However, a GRU does not consider adding these values into the formula (Eq. 3.9) and (Eq. 3.10). The other difference is that instead of changing the current hidden layer h as in the LSTM method, the input x and the previous layer $h_{(t-1)}$ modify the update gate and the reset gate values in the GRU method. Then the current layer is updated accordingly by z and r (Eq. 3.12) (E. Choi et al. 2016).

3.1.3 Autoencoders

Autoencoders, introduced by Hinton Rumelhart and McClelland (1986) are unsupervised learning algorithms as they are used to learn representations of the inputs from unlabelled data (Chalapathy, Menon and S. Chawla 2018). Autoencoders have been used effectively in tasks that involve dimensions reduction and single-class learning (anomaly detection) (Van Der Maaten, Postma and Van den Herik 2009). They can learn the correlations between the input features by transforming them into a latent space with new encoded dimensions. Decoding a latent space back into an input can help the autoencoder to learn hidden features, correlations, and patterns in the data.

Autoencoder networks have been demonstrated as having the capability to address tasks that involve anomaly detection. Unlike other deep learning methods, autoencoders are trained using unsupervised algorithms that can learn from single-class data, attempting to represent its input x as reconstructions r. The autoencoder networks consist of an encoder function $h = f(x)$, followed by a decoder function that generates reconstructions using the decoder function $r = g(h)$ (LeCun, Bengio and Hinton 2015; An and S. Cho 2015). A stacked autoencoder (SA) uses more than one encoding and/or decoding layer (function). The output of each layer is used as the input for the successive layer.

The autoencoder's minimise the errors (known as reconstruction errors) between the input values x and the reconstructions using loss functions L (Eq. 3.13) for the learning process (Goodfellow et al. 2016).

$$L(x, g(f(x))) \tag{3.13}$$

The *identity function* problem occurs when the autoencoder network cannot extract important features from the input and memorises the data instead of learning the useful and discriminative patterns. This can be avoided by limiting the number of units in the hidden layers to be fewer than the number of input units. This means mapping the input into a lower dimensional space

(new features). This can help the network to learn correlations between the input features. These encoded-aggregated features are called latent-features. In contrast, having a greater number of hidden units can also be helpful for an autoencoder network, especially when imposing sparsity on the hidden units.

Another important extension of the autoencoder architecture is the Denoising Autoencoder (DAE). DAE tends to force the hidden units to extract features from a corrupted version $\tilde{x}$ of the original input x. Decoding a corrupted input can help the network to learn to extract important features and avoid the identity function problem by undoing the corruption (Steck 2020). The DAE tries to minimise the reconstruction error using $\tilde{x}$ (Eq. 3.14).

$$L(x, g(f(\tilde{x}))) \tag{3.14}$$

In our work, we investigate the performance of a Stacked Denoising Autoencoders (SDA) model using sequences of patient observations as input $x : x_1, x_2, ...x_n$. We use the mean squared error (MSE) function (Eq. 3.15) to calculate the reconstruction error after fitting the model with test data.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - r)^2 \tag{3.15}$$

3.2 Conventional Machine Learning Approaches

This section gives a brief overview of the main conventional machine learning approaches employed in this thesis. Since the main objective of this work is to investigate deep learning models, we use common machine learning models (briefly described below) for purpose of comparison.

Support Vector Machine

Support Vector Machine (SVM) was introduced in 1998 (Vapnik 2013). SVM can be used to solve both classification and regression problems. It uses the training feature space to decide on the separation boundaries (hyperplane) that best divides the training dataset into regions (one for each class). The points closest to separation hyperplanes are called 'support vectors'. SVMs often use kernels to help enhance class separation by mapping the training features into a larger space with an increased number of dimensions (Noble 2006; Vapnik 2013).

Random Forest

Random Forest (RF) is a very common algorithm used for classification. It combines several decision trees that are generated during the training process. Each decision tree is trained using a random subset of the training dataset. The final classification is then based on the majority vote of all the generated decision trees (Breiman 2001).

Logistic regression

Logistic Regression (LR) is commonly used to solve binary classification problems. It calculates the odds ratio of the variables and is similar to multiple linear regression but uses a binomial distribution of the dependent variable (i.e. more than one). It includes a logit function that handles different types of relationships between the dependent and independent variables (Rawlings, Pantula and Dickey 2001; Sperandei 2014).

In this thesis, we have chosen the above models for the purpose of comparison (refer to the comparative models used in chapter 6 section 6.2, chapter 7 section 7.3.1, and chapter 8 section 8.2) because they have been shown to be frequently/commonly used in recent studies (Gray et al. 2012; Stanley Xu et al. 2014; Awad et al. 2017; Dwivedi 2018; Faisal et al. 2020). Details about the models used in recent studies will be presented/discussed in the next chapter.

3.3 Measures Used for the Evaluation of Model Performance

We have used a number of different measures for evaluating the performance of the models employed in our studies. The measures have been selected for their wide use in machine learning predictive models (such as AUC-ROC and F1 (Moon et al. 2020; Mandrekar 2010)) and suitability for research in the medical domain, having them used in the studies (to be reviewed in Chapter 4) that are strongly related to this work (Lipton et al. 2015; Miotto et al. 2016; E. Choi et al. 2016; Wells, Lenoir et al. 2018). Other measures (such as precision and recall) are reported as these give more understanding of the performance of the machine learning models. Their suitability for use with balanced/imbalanced data (Schütze, Manning and Raghavan 2008) and the enabling of comparison with other studies form other reasons for selecting those measures. The confusion matrix, explained in Figure 3.4, is mainly used for evaluating the performance of predictive models using variety of meaningful measure.

Data class	Classified as positive	Classified as negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

$$\begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}$$

Figure 3.4: The confusion matrix used for predictive models evaluation.

In medical predictive classification (disease diagnosis specifically) the cost of misclassifying positive patients will very often have more impact on patients than misclassifying negative cases (E. Choi et al. 2016). Hence, evaluating the performance of such classifiers would give more attentions to the measures that gives more weights to true positive predictions (such as sensitivity (Recall), precision, F1, Area Under the Receiver Operating Characteristic (AUROC)) and ultimately the overall accuracy. Below are the details of the measures used in the studies of this research.

The total accuracy provides a general indicator for the overall performance of the classifier. It measures the overall accuracy (giving the same weighting for both negative and positive predictions) by dividing the total number of correctly classified cases by the total number of cases. It is calculated using Equation 3.16

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.16}$$

The recall measures how well the classifier predicts positive cases using Equation 3.17:

$$Recall = \frac{TP}{TP + FN} \tag{3.17}$$

Precision is used to calculate the ratio of positive predicted cases that are in fact true positive cases using Equation 3.18:

$$Precision = \frac{TP}{TP + FP} \tag{3.18}$$

The F1 measure can be considered a trade off between precision and recall. It is calculated using Equation 3.19:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.19}$$

While the F1 score gives same weighting to the classes, the F1-Weighted score gives weighting to the classes by number of samples for the classes. The macro measure for F1, recall and precision gives more weight to the small classes in the test collection (imbalanced) data (Schütze, Manning and Raghavan 2008) .

The Receiver Operating Curve (ROC) plots the performance of a given classifier at different thresholds using true positive and false positive rates. The measure AUR-ROC is an aggregation of the classifier performance using all possible thresholds (Austin and Steyerberg 2012). It calculates the probability of classifier with higher ranking of randomly selected positive cases than randomly chosen negative cases.

Epilogue

In this chapter,we have presented the machine learning approaches employed in this thesis. The employed machine learning technologies (including deep and conventional approaches) were introduced. In the next chapter, we review the existing predictive methods and dataset used in literature in the fields of Type-2 Diabetes Mellitus (T2DM), HbA1c elevation levels, and in-hospital mortality risk predictions.

Chapter 4

Related Work

Prologue

Machine learning models have shown powerful capabilities for analysing and understanding complex data across a wide variety of applications. This chapter is aimed at reviewing and discussing the current predictive methods and datasets used in the literature for Type-2 Diabetes Mellitus (T2DM), HbA1c levels, and for in-hospital mortality risk prediction challenges. Since there are a large number of studies related to diabetes, these studies are organised by the predictive technique used, neural and non-neural (conventional) models. The outcome from any empirical study such as these reviewed in this chapter, depends both upon the technique used and also the dataset employed. After reviewing the studies, discussion of the techniques and datasets used for diabetes and HbA1c prediction will be presented in subsection 4.1.4 and in 4.2.1 for those used for mortality risk prediction.

4.1 Machine Learning in T2DM and HbA1c Predictions

Although this work focuses on the prediction of HbA1c, we also consider it important to survey recent studies on diabetes diagnosis and pattern recognition using machine learning as well as

studies that employed machine learning for HbA1c prediction. Therefore, this section begins by reviewing studies related to diabetes followed by those related to HbA1c.

4.1.1 Conventional Machine Learning Models in T2DM Diagnosis

Machine learning has long been applied to the analysis of medical datasets through a variety of algorithms (Kononenko 2001). Therefore, a relatively large number of studies exist on diabetes diagnosis using conventional machine learning models such as decision trees (RF) (Breiman 2001), Naive Bayes Classifiers (NB) (Domingos and Pazzani 1997), and Support Vector Machine (SVM) (Suykens and Vandewalle 1999). We examine each of these in turn below.

A study by J. Wu et al. (2009) used a different version of SVM called the LapSVM algorithm; this depends on a manifold regularisation technique to help classify semi-supervised data. The authors compared the LapSVM's performance on classifying semi-supervised (partially labelled) and supervised (fully labelled) datasets. The LapSVM algorithm was applied to the Pima Indian Diabetes Databas (PIDD) (Lichman 2013). The authors modified the dataset so that it was semi-supervised by randomly removing the labels of some instances in the dataset. The study concluded that the semi-supervised method helped to increase the total accuracy of the classifier. Besides, Wu et al., Polat, Güneş and Arslan (2008) and the study by Aishwarya, Gayathri et al. (2013) all had the same objective; these studies provided useful techniques for data pre-processing (i.e. features selection and changing the dataset to semi-supervised) to enhance the accuracy of SVM models.

Hippisley-Cox et al. (2009) applied the Cox proportional hazards model to the QResearch database (version 19) to predict patients' risk of developing diabetes over the next ten years. The Cox proportional hazards model is a statistical regression model commonly used in medical research to identify the association between the predictors and survival time by measuring the hazard rate (risk of failure) (Lin and Wei 1989). The QResearch database (version 19) is a large EHR dataset collected from 551 general practices in England and Wales. The data from 355 general practices were randomly chosen for training the model and 177 for validation. The

overall size of the dataset used was 3,773,585 samples, of which 115,616 patients were diagnosed with diabetes in the follow-up period (10 years). The study concluded that the used regression model is suitable for clinical use and self-assessment.

A study by Mani et al. (2012) explored the diabetes prediction using several machine learning models — SVM, decision tree, Gaussian Naïve Bayes and LR — to predict the onset of diabetes in the next 6 or 12 months. Mani et al. used an EHR dataset containing 2280 samples, of which 228 were patients with diabetes. The results showed that T2DM prediction using an EHR dataset is feasible. This study was among the first to employ machine learning to the prediction of T2DM using EHR data.

The PIDD along with another dataset from a local hospital (that used different classification techniques) were used by Velu and Kashwan (2013). The authors investigated three techniques: (i) the Expectation Maximisation (EM) algorithm, which uses an iterative method to find the maximum likelihood explained by Dempster, Laird and Rubin (1977); (ii) h-means+ clustering, which is an improved version of the k-mean algorithm; and (iii) the Genetic Algorithm (GA) algorithm to create clusters of similar diabetes symptoms. The results revealed that the h-means+ algorithm performed better compared to EM and GA. The results also revealed that EM performed slightly better when employed with fewer dimensions of data. However, the main goal of this study was to create clusters to discover patterns using the above likelihood approaches.

A study by Kaur and Chhabra (2014) applied an improved decision tree classification algorithm (referred to as C4.5 J48) developed by Ross Quinlan (1993), on the PIDD dataset. The modified decision tree algorithm had three key features: (i) an improved algorithm was used to help the processing of discrete and continuous attributes; (ii) missing data was handled by this algorithm; and (iii) it used data pruning methods to remove noisy examples from the data. The classifier showed a considerable performance improvement compared to the existing standard J48 algorithm. It is also important to note that the authors used overall (total) accuracy and random accuracy measures only for evaluating the performance of the modified classifier. In

addition, it was not clear in this study how many instances were excluded after applying the pruning methods.

The study by Kaur and Chhabra (2014) was followed by similar studies by Anderson et al. (2016) and Rallapalli and Suryakanthi (2016). The authors of both studies analysed the performance of two commonly used conventional models. First, Anderson et al. (2016) employed the Multiple Logistic Regression (MLR) and RF models. The dataset contained 9,948 unique patient records, of which only 1,805 patients had T2DM. Meanwhile, Rallapalli and Suryakanthi (2016) also employed a RF model along with a Naive Bayes and a variety of regression models. Both studies revealed that the EHR dataset can be successfully used for the detection of T2DM.

4.1.2 Neural Networks Models in T2DM Diagnosis

Artificial Neural Networks (ANN) has shown powerful capabilities for analysing and understanding complex clinical data in a variety of medical applications (Hill et al. 1994). Specifically, diabetes diagnosis and pattern recognition were key medical applications that attracted the attention of researchers for the use with neural networks models (Al-Shayea 2011). Therefore, using neural networks for the diagnosis of diabetes has been the subject of much discussion and many studies have used this approach.

Venkatesan and Anitha (2006) studied applying a MLP model (a type of Feed Forward Neural Networks (FFNN) that uses mathematical functions to map input to output values) together with a Radial Basis Function (RBF) network to diagnose diabetes. RBF is another type of FFNN that uses a function (e.g. Gaussian) for each neuron in the hidden layer (RBF conceptually is similar to KNN, which means that the predicted class is likely to be the same as the closest classes). The authors compared the results with a LR method. Two different datasets with 1200 and 600 examples were used in this work. With regard to sensitivity, specificity, and overall accuracy measures, the experiment showed that MLP and RBF Feed Forward Neural Networks (FFNN) models performed significantly better than the LR model.

Two BPN networks with different architectures — one a single hidden layer [6-10-1] architecture and the other one was double hidden layer [6-14-14-1] — were used by Dey et al. (2008). The authors used a dataset of 530 samples from the Sikkim Manipal Institute of Medical Sciences hospital to train and test the ANN models. However, the ANN models did not always achieve better results than the conventional models.

Different neural network models have been used in conjunction with other data pre-processing techniques to address the impact of data pre-processing on the performance of the neural network models. The use of a Generalised Discriminant Analysis (GDA) prior to Least Square Support Vector Machine (LS-SVM) was proposed by Polat, Güneş and Arslan (2008). GDA is a method that transforms spaces into a high-dimensional feature space for nonlinear classification based on kernel functions. While LS-SVM is similar to SVM, the LS-SVM uses linear rather than quadratic equations. Using the PIDD dataset, GDA was used as a data pre-processing method to reduce the number of features prior to the use of LS-SVM. The results showed that the GDA-LS-SVM approach outperformed the standard LS-SVM algorithm in the diagnosis of diabetes task.

Another technique based on deep neural networks was applied by H. Temurtas, Yumusak and F. Temurtas (2009), who compared the result of applying two neural network models with the results of previous studies using the PIDD dataset. The first was the MLP model using Levenberg–Marquardt (LM) for quicker convergence and better performance. The second was the Probabilistic Neural Network (PNN) model that uses a statistical method called 'Bayesian classifiers', which aids classification. The results also showed that MLP performed better with regard to accuracy when using 10-fold cross-validation compared to previous studies. The author also reported that the classification accuracy obtained using MLP with LM was always better than the corresponding classifiers even when a conventional validation technique was used in comparison. Although LM is considered a fast algorithm, in some cases, LM can cause memorisation problems, which may impact the generalisation of the neural network, and hence affecting performance (Meng et al. 2013).

A study by Karegowda, Manjunath and Jayaram (2011) used the integrated GA algorithm (a metaheuristic algorithm for optimisation) together with BPN. Propagation involved a process of adjusting the weights after computing the error for each iteration. This study aimed to use a GA to determine the optimal parameters for the neural network — such as the number of hidden layers and input variables. It is also important to highlight that the authors used Correlation-based Feature Selection (CFS) — which uses a heuristic method to evaluate the features (M. A. Hall 2000), along with GA in the features selection step. Using the PIDD dataset, the result showed a significant increase in the overall accuracy of the GA-BPN approach. M. Pradhan and Sahu (2011) also proposed a similar approach to Karegowda el al. with slight differences. Pradhan and Sahu combined the GA with a single hidden layer of BPN without taking into account reducing the size of the features.

Jayalakshmi and Santhakumaran (2011) applied several normalisation methods used to re-scale the data in specific ranges on the PIDD dataset before using the BPN. The results revealed that the performance of the neural network varied depending on the re-scaling approaches used. The Statistical Column re-scaling method showed a significant improvement in the performance of the neural network.

A new ANN approach called Extreme Learning Machine (ELM) has shown very promising results for regression and classification problems. ELM combines the LS-SVM (Suykens and Vandewalle 1999) and Proximal Support Vector Machine (PSVM) (Fung and Mangasarian 2005; Huang et al. 2012), and uses a Single Layer Feed Forward Neural Network architecture. It randomly chooses the input weights without back tuning (Huang et al. 2012; Priyadarshini, Dash and Mishra 2014). Huang et al. (2012) reported that ELM achieves state-of-the-art or better generalisation performance for binary classes and regression classification problems. Besides, the authors showed that ELM achieves much better generalisation in multi-class classification problems. Priyadarshini, Dash and Mishra (2014) applied ELM with the PIDD dataset and compared it with Back Propagation Neural Network (BPM). ELM achieved better accuracy and was faster than BPN. Similarly, Pangaribuan et al. (2014) employed ELM with a simpler

structure (only one hidden layer); this model showed promising performance as it achieved a faster training time compared to BPN.

In contrast to the study by H. Temurtas, Yumusak and F. Temurtas (2009), a more recent study by Meng et al. (2013) applied three classification models to a different dataset of 1487 examples with 12 attributes (which is double the size of the PIDD dataset used in most of the other studies). The study used the LR, modified Decision Tree (an improved version of Decision Tree), and MLP classification models. Although the LR and modified Decision Tree algorithms performed better in terms of total accuracy in this study, the MLP network showed a promising total accuracy besides producing the highest sensitivity (recall) accuracy.

Motka et al. (2013) showed a significant increase in the performance of ANN and Artificial Neural Fuzzy Interference Systems (ANFIS), which integrate fuzzy logic with a neural network, after using Principal Component Analysis (PCA). The PCA uses a non-parametric method to reduce feature dimensions. It was used with the PIDD dataset to reduce the dimensionality. The author concluded that the performance of ANN can be enhanced when combined with other techniques such as PCA. The same approach was used in (Polat and Güneş 2007); here, PCA suggested reducing the PIDD dimensions from 8 to only 4. The study also showed that PCA-ANFIS can help with the diagnosis of diabetes.

A study by Sarwar and Sharma (2014) addressed the use of neural networks rather than conventional classifiers. The authors showed that even with a small dataset size for training and testing, the accuracy of ANN outperforms the Naive Bayes Network — a probabilistic classifier that uses Bayes theorem — and K-Nearest Neighbours (KNN), which is an instance-based classification based on a number of close training examples. The authors of this study collected a dataset containing only 500 samples with 10 attributes. The ANN model performed better than the rest of the models used in this study. However, this study suggests that ANN can perform well in small datasets; this contrasts with the results of other studies such as (Raudys, Jain et al. 1991) and (Mazurowski et al. 2008), which assumed that the more examples the better for future classification queries.

A study by Lipton et al. (2015) — who claimed that this study was the first study of its kind — applied LSTM-RNN to health care datasets. The authors used LSTM-RNN on an Intensive Care Unit (ICU) dataset to predict health conditions, events, and treatment responses. They observed that while a linear-increase weight technique provided improvements in Yue-Hei Ng et al. (2015) and Dai and Q. V. Le (2015) studies, this technique decreased performance in this study. However, the LSTM-RNN in this study showed promising accuracy performance using AUC, precision, recall, and F1 (harmonious weight between precision and recall) measures to evaluate the performance of the LSTM-RNN model.

Rau et al. (2016) compared ANN and LR liver cancer prediction models for those diagnosed with diabetes six years earlier. The authors used a dataset of 2060 examples with a total of 10 attributes (risk factors). Although this study revealed the ANN prediction model achieved good performance, only a small and imbalanced dataset was used — with a class distribution of 25% to 75%.

A study by Miotto et al. (2016) proposed a framework to extract general features using a deep-learning approach with a large-scale medical dataset. The framework used independently trained layers of Stacked Denoising Autoencoders (SDA) (detailed in Chapter 3) to represent features from structured, semi-structured, and unstructured data. The extracted features were then used with RF to predict a variety of diseases. However, in this study, the deep learning model was not used for disease prediction nor pattern recognition; instead, it was applied to find general features from the large dataset.

Another study by E. Choi et al. (2016), applied GRU-RNN to larger and longitudinal patient datasets extracted from EHR systems to make predictions on disease diagnosis and medication categories for next visits. The authors explored using four different RNN architectures with different numbers of layers and initialisation approaches. The results showed that RNN achieved 79.5% recall accuracy, which outperformed the comparative neural network models used.

Esteban et al. (2017) explored different conventional and deep machine learning models using calculated variables extracted from an EHR dataset collected from Argentina. The population

size was 2,463 samples with unbalanced class distribution. The models were trained using an artificially balanced dataset using oversampling techniques. The features selected as input in all models were calculated variables such as number of HbA1c tests less than 6.5% and number of HbA1c tests higher than or equal to 6.5%. Six calculated features were employed in total. The results revealed that the MLP model showed promising classification performance using EHR data for the screening of T2DM.

A recent study by Zou et al. (2018), applied decision tree, RF, and neural network classifiers to predict the diabetes diagnosis using 14 physical examination variables. The authors used a large dataset of 220,680 examples. Due to the data imbalance, the models were trained using under-sampled data. The authors also investigated applying PCA and minimum Redundancy Maximum Relevance (mRMR) to reduce the dimensionality. The results showed that RF achieved best performance using all variables. The obtained results confirmed that reducing the dimensionality of the data using PCA and mRMR did not help improve the performance of the models used.

4.1.3 Machine Learning in the Prediction of HbA1c Levels

Studying the trend of HbA1c levels in patients is an interesting area of research in healthcare informatics. Some studies have investigated the correlation between HbA1c levels and clinical variables. McCarter, Hempe and Chalew (2006) studied the association between T2DM patients' HbA1c levels and clinical variables. Conversely, Nathan et al. (2008) were able to demonstrate promising results for calculating clinical variables — specifically, average glucose levels — from HbA1c levels. Other work by Rose and Ketchell (2003) showed a correlation between mean blood glucose level and HbA1c level. The correlation coefficients were found to be between 0.71 to 0.86. However, the results can be significantly affected by the time of day (before or after meals) at which blood glucose levels were measured.

Other studies used statistical models based on data collected from diabetic patients. Stanley Xu et al. (2014) used a linear regression model for missing HbA1c data imputation. Their model calculates HbA1c levels for patient records with missing HbA1c values as continuous and

categorical values, and used four features extracted from an EHR system: (i) RBS, (ii) fasting blood sugar (FBS), along with (iii) age and (v) gender, as predictors to calculate HbA1c levels. The authors used samples of newly identified patients with diabetes.

Kazemi et al. (2014) added complications — such as retinopathy — to the clinical variables to analyse the trends in HbA1c levels. The objective of the study was to identify variables that increased and those that decreased the trends of HbA1c using a linear regression model that combined the longitudinal readings in the input of the model. Ngufor et al. (2019) used a machine-learning-based framework to predict the longitudinal changes in HbA1c for adult patients with diabetes. Koopman et al. (2017) used linear regression models to predict HbA1c levels after six years for non-diabetic patients using different populations.

A study by S. B. Choi et al. (2014) developed models for patients with pre-diabetes screening. The authors used a survey dataset to train, validate, and test the developed models. Seven variables were selected as the optimal variables (age, BMI, hypertension, gender, alcohol intake, waist circumference, and family history) to train the models. These variables were selected by calculating the variables' correlation with the outcome using backward logistic regression. The outcome, or the classification class, was calculated based on the levels of fasting plasma glucose (FPG) (between 100–125 mg/dL or 5.6–6.9 mmol/L), impaired glucose tolerance (IGT) (between 140–199 mg/dL or 7.8–11.0 mmol/L) or HbA1c (between 5.7–6.4%). The authors designed two models, SVM and ANN, to screen the patients for pre-diabetes. The SVM model achieved better results than the results obtained using the neural network model.

A recent study by Wells, Lenoir et al. (2018) was the first to focus on predicting current HbA1c elevation levels for non-diabetic patients using an EHR dataset. Multiple logistic regression was employed to create an equation that calculates the probability of having an elevated HbA1c level (>=5.7). The dataset was extracted from an EHR system used for patients in the USA. The authors used eight independent variables (age, body mass index (BMI), random glucose (RBS), race, serum non–high-density lipoprotein (non-HDL), serum total cholesterol, estimated glomerular filtration rate (eGFR), and smoking status) fitted to the model using Restricted

Cubic Splines (RCS) with 3-knots to formulate the final equation. The calculator performance created by Wells et al. was compared to the models of Baan et al. (1999) and Griffin et al. (2000) (However, the models by Baan and Griffin aimed at predicting the onset of patients' diabetes rather than predicting HbA1c levels for non-diabetic patients).

4.1.4 Discussion of Datasets and Predictive Methods Used for Diabetes Diagnosis and HbA1c Level Predictions

In this literature, a variety of studies have investigated diabetes diagnosis and HbA1c predictions using machine learning models. Studies related to predictive models for T2DM diagnosis and HbA1c level predictions in the literature review of this thesis will be discussed from two aspects: (i) the uncovered gaps in the predictive models employed and (ii) the issues in the datasets used. The characteristics by which the datasets will be discussed/reviewed are: dataset availability, type (driven from EHR or non-EHR data), number of variables (also referred as "features" or "predictors"), size (number of samples/records used to train/test the models), class balance (distribution of the classes), and date of the dataset (as shown in Table 4.1).

The size of the dataset used to train and test the models tends to attribute to the success of deep learning models (C. Sun et al. 2017). Also, using balanced datasets eases the difficulties inherent in classification learnability for machine learning models (Galar et al. 2011). The imbalance occurs when the presence of one class in the dataset exceeds its counterpart class. While what constitutes data imbalance ratio between the classes of a dataset is debatable, the imbalanced datasets are an issue for machine learning algorithms, especially in case of disease classification (Batista, Prati and Monard 2004). Training predictive models using imbalanced datasets tends to work poorly especially when predicting the minority class (Rahman and Davis 2013). At the same time, solving this problem by artificially balancing medical datasets does not significantly impact the predictive efficiency of the machine learning models (Provost 2000; Gu et al. 2008). Therefore, using originally imbalanced datasets is, to date, a challenging obstacle for training machine learning models.

T2DM Diagnosis

As illustrated in Table 4.1 (which lists the main characteristics of the datasets used in the above studies sorted by predictive method and date), it is very noticeable that the Pima Indian Dataset (PIDD) (Lichman 2013) has been intensively investigated for diabetes diagnosis prediction using a variety of machine learning approaches. PIDD was publicly available from the University of California, Irvine (UCI) repository of machine learning databases, and it contains 786 records with 8 features. It is important to note that the PIDD dataset was collected from a female-only population.

As shown in the table, the majority of the datasets are not publicly available. However, these datasets suffer from several problems. The main issue with the non-EHR datasets (except DM15) — DM1, DM6, DM7, DM8, DM9, and DM10 — is the small size of the samples used to train and test the predictive models. The studies that used EHR datasets — DM2, DM3, DM4, DM5, DM12, DM13, and DM14 — use a larger sample size; however, both the non-EHR and EHR datasets were found to be extremely imbalanced, except for datasets DM9 and DM10.

Table 4.2 lists the models employed in the above studies (sorted by study date). The table shows that a relatively large number of studies investigated T2DM diagnosis classification. However, the table also shows that few studies have investigated applying deep learning models for this task. Furthermore, no studies have yet investigated the temporal effect of time series data (vital signs or lab test results) for T2DM diagnosis prediction.

A few recent studies have investigated the temporal effect of time-series data in EHR systems. Those studies used RNN models together with general clinical time series datasets for multi-disease diagnosis classification (Lipton et al. 2015; E. Choi et al. 2016). However, the time series datasets and the models employed in these studies were not used for the purpose of diabetes diagnosis specifically.

Lipton et al. (2015) proposed the first model that applied LSTM on a clinical dataset. The authors used LSTM on a Children's Intensive Care Unit (ICU) dataset to predict multiple diseases

Table 4.1: Details about the datasets used in the literature for T2DM diagnosis prediction.

Studies	Avail[a]	Type	No. Var[b]	No. Rec[c]	Comments
Dataset(DM1): PIDD used by Kaur and Chhabra (2014), Velu and Kashwan (2013), Polat, Güneş and Arslan (2008), Aishwarya, Gayathri et al. (2013) , H. Temurtas, Yumusak and F. Temurtas (2009), Karegowda, Manjunath and Jayaram (2011), M. Pradhan and Sahu (2011), Motka et al. (2013), Jayalakshmi and Santhakumaran (2011), Huang et al. (2012), Priyadarshini, Dash and Mishra (2014), Dwivedi (2018), Vijayan and Anjali (2015)	✓	Non-EHR	9	768	- Class distribution: Diabetic 35% vs Normal 65% - Dataset date: 1990 - Other comment: Female examples only
Dataset(DM2): QResearch database used by Hippisley-Cox et al. (2009)	✗	EHR	9	3,773,585	- Class distribution: Diabetic 3% vs Normal 97% - Dataset date: NM
Dataset(DM3): Synthetic Derivative (SD) used by Mani et al. (2012)	NM	EHR	12	2,280	- Class distribution: Diabetic 10% vs Normal 90% - Dataset date: NM
Dataset(DM4): used by Rallapalli and Suryakanthi (2016)	✗	EHR	NA	NM	- Class distribution: NM - Dataset date: NM
Dataset(DM5): used by Vijayan and Anjali (2015), Anderson et al. (2016)	✗	EHR	NA[e]	9,948	- Class distribution: Diabetic 18% vs Normal 82% - Dataset date: 2012
Dataset(DM6): used by Venkatesan and Anitha (2006)	✗	Non-EHR	9	1,800	- Class distribution: Diabetic 55% vs Normal 45% - Dataset date: 1998
Dataset(DM7): used by Sarwar and Sharma (2014)	✗	Non-EHR	10	500	- Class distribution: NM[d] - Dataset date: 2012
Dataset(DM8): used by Rau et al. (2016)	✗	Non-EHR	10	2,060	- Class distribution: Diabetic 100% vs Normal 0% - Dataset Date: 2009 - Other comment: Used for predicting liver cancer in diabetic patients
Dataset(DM9): used by Dey et al. (2008)	✗	Non-EHR	6	530	- Class distribution: Diabetic 47% vs Normal 53% - Dataset date: NM
Dataset(DM10): used by Meng et al. (2013)	✗	Non-EHR	12	1,487	- Class distribution: Diabetic 49% vs Normal 51% - Dataset date: 2007
Dataset(DM11): used by Lipton et al. (2015)	✗	EHR	13	10,401	- Class distribution: NA - Dataset Date: 2015 - Other comment: Children samples only from Intensive Care Unit (ICU) data.
Dataset(DM12): used by E. Choi et al. (2016)	✗	EHR	NA	260,000	- Class distribution: NA - Dataset date: NM
Dataset(DM13): used by Miotto et al. (2016)	✗	EHR	NA	704,857	- Class distribution: NA - Dataset date: 2014 - Other comment: Used for deep feature selection purpose.
Dataset(DM14): used by Esteban et al. (2017),	✗	EHR	NA	2,463	- Class distribution: NM - Dataset date: 2015
Dataset(DM15): used by Zou et al. (2018)	✗	Non-EHR	14	220,680	- Class distribution: Diabetic 69% vs Normal 31% - Dataset date: NM

- [a]Avail: Availability. [b]No. Var: Number of variables. [c]No. Rec: Number of records. [d]NM: Not mentioned. [e]NA: Not available.

diagnosis (such as asthma, hypertension and anaemia) using 13 lab test results. The LSTM model was built to classify 128 diseases with competitive accuracy. The other study (E. Choi et al. 2016), applied GRU to larger and longitudinal patient data extracted from general patients' clinical records. Similar to Lipton's study, the aim of this study was mainly to predict disease diagnosis. However, the features used in this study are different in type than the ones used in Lipton's study. The authors did not make use of patient observation records (vital signs or lab test results). Instead, they used previous patient diagnoses as input to predict future diseases. However, it is unclear how many and what diseases were examined to evaluate the model.

Machine Learning for HbA1c Levels Prediction

Several studies investigated the use of machine learning to predict the risk of diabetes and pre-diabetes. The studies by Robinson, Agarwal and Nerenberg (2011), Hische et al. (2010), DuBose et al. (2012) and Xin et al. (2010) investigated predicting patients with impaired glucose (diabetes and pre-diabetes) using Oral Glucose Tolerance Test (OGTT) and/or Fasting Blood Sugar (FBS). Those studies used logistic regression and/or decision trees algorithms using non-EHR dataset.

This thesis focuses on studies that have employed the predictive models to help the early identification of patients with risk of diabetes or pre-diabetes using elevated levels of HbA1c. Therefore, studies that did not used HbA1c to assess patients with risk of diabetes or pre-diabetes have been excluded. Besides, the studies that used data collected from diabetic patients only have been also excluded from the discussion in the thesis (such as (Stanley Xu et al. 2014) and (Kazemi et al. 2014)). As shown in Table 4.3, those studies used samples from a population of patients with diabetes. Studies that investigated the correlation between HbA1c levels and clinical variables are also outside the scope of the present thesis.

Table 4.3 lists the characteristics of the datasets used in studies aimed at predicting the levels of HbA1c for patients. Gray et al. (2012) used logistic regression to identify patients with diabetes and pre-diabetes. The study used both the OGTT or HbA1c to identify patients with impaired glucose regulation. For this study, the data was collected from the primary care records and by

Table 4.2: Summary of predictive models used in the literature for T2DM diagnosis.

Study	**Method**	**Dataset**	**Year**
Venkatesan and Anitha (2006)	MLP	Dataset(DM6)	2006
Polat, Güneş and Arslan (2008)	SVM	Dataset(DM1)	2008
Dey et al. (2008)	BPN	Dataset(DM9)	2008
Hippisley-Cox et al. (2009)	Cox proportional hazards model	Dataset(DM2)	2009
H. Temurtas, Yumusak and F. Temurtas (2009)	MLP and Probabilistic Neural Network (PNN)	Dataset(DM1)	2009
Karegowda, Manjunath and Jayaram (2011)	Genetic Optimisation Algorithm (GA) with BPN	Dataset(DM1)	2011
M. Pradhan and Sahu (2011)	Genetic Genetic Optimisation Algorithm (GA) with BPN	Dataset(DM1)	2011
Jayalakshmi and Santhakumaran (2011)	BPN	Dataset(DM1)	2011
Huang et al. (2012)	ELM	Dataset(DM1)	2012
Mani et al. (2012)	SVM, decision tree, Gaussian Naïve Bayes and LR	Dataset(DM3)	2012
Velu and Kashwan (2013)	EM, h-means+ clustering and GA	Dataset(DM1)	2013
Meng et al. (2013)	LR, modified Decision Tree and MLP	Dataset(DM10)	2013
Motka et al. (2013)	Artificial Neural Fuzzy Interference Systems (ANFIS)	Dataset(DM1)	2013
Aishwarya, Gayathri et al. (2013)	SVM	Dataset(DM1)	2013
Kaur and Chhabra (2014)	Customised decision tree (C4.5 J48)	Dataset(DM1)	2014
Priyadarshini, Dash and Mishra (2014)	ELM	Dataset(DM1)	2014
Sarwar and Sharma (2014)	ANN, Naive Bayes Network and KNN	Dataset(DM7)	2014
Lipton et al. (2015)	Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN)	Dataset(DM11)	2015
Rau et al. (2016)	ANN and LR	Dataset(DM8)	2016
E. Choi et al. (2016)	Gated Recurrent Units with Recurrent Neural Networks (GRU-RRN)	Dataset(DM12)	2016
Miotto et al. (2016)	SDA with RF	Dataset(DM13)	2016
Anderson et al. (2016)	MLR and RF	Dataset(DM5)	2016
Rallapalli and Suryakanthi (2016)	Logistic Naive bayes, LR and RF	Dataset(DM4))	2016
Esteban et al. (2017)	Several rule-based statistical, stacked generalisation models	Dataset(DM14)	2017
Dwivedi (2018)	ANN, Classification tree, Naïve Bayes, LR, SVM and KNN	Dataset(DM1)	2018
Zou et al. (2018)	Decision tree, RF, ANN	Dataset(DM15)	2018

running a questionnaire. The study by Handlos et al. (2013) used Multiple Logistic Regression (MLR) with 14 variables (predictors) collected from the clinical data and a questionnaire as well. The study by S. B. Choi et al. (2014) aimed to identify patients with pre-diabetes using a survey dataset.

To the best of our knowledge, the only study that applied a predictive model to predict the elevated levels of HbA1c using an EHR dataset was by Wells, Lenoir et al. (2018). However, Wells et al.'s study used a simple statistical model using MLR to calculate the probability of HbA1c levels. Table 4.4 shows the models employed in the related work studies.

Table 4.3: Details about the datasets used in the literature for HbA1c levels predictions.

Studies	**Avail**[a]	**Type**	**No. Var**[b]	**No. Rec**[c]	**Comments**
Dataset(A1C1): ADDITION-Leicester used by Gray et al. (2012)	✗	Primary care and questionnaire	14	6,390	- Class distribution: Normal 71.1% vs 29.9% Abnormal (class distribution for test data: Normal 64.6% vs 35.4% Abnormal) - Dataset date: NM
Dataset(A1C2): used by Handlos et al. (2013)	✗	Clinical data and questionnaire	8	6,588	- Class distribution: Normal 82.2% vs 7.8% Diabetic and 10.0% Pre-diabetes - Dataset date: 2010-2011
Dataset(A1C3): SUPREME-DM dataset used by Stanley Xu et al. (2014)	✗	EHR	4	62,458	- Only incident samples of diabetes - Dataset date: 2005 - 2010
Dataset(A1C4): used Kazemi et al. (2014)	✗	NM[d]	NM	3,440	- Only incident samples of diabetes - Dataset date: 2000 - 2012
Dataset(A1C5): used by S. B. Choi et al. (2014)	✗	Non-EHR (Survey dataset)	7	9,251	- Class distribution: Normal 79% vs 21% Pre-diabetes (class distribution for the test dataset was not mentioned) - Dataset date: 2010-2011
Dataset(A1C6): used by Wells, Lenoir et al. (2018)	✗	EHR	8	22,635	- Class distribution: Normal 74% vs 26% Pre-diabetes - Dataset date: 2012-2016

- [a]Avail: Availability. [b]No. Var: Number of variables. [c]No. Rec: Number of records. [d]NM: Not mentioned.

As shown in Table 4.3, the datasets used to train and validate the models developed in Choi et al.'s and Wells et al.'s studies were relatively small datasets (9k and 22k, respectively). In addition, the training dataset used by Choi et al. was significantly imbalanced: 79% of the sample were patients in normal health while 21% were patients with diabetes. Similarly, the dataset used by Wells et al. to train and test their model was also imbalanced with 74% of the samples having normal HbA1c levels (<5.7%) and only 26% of the samples having elevated HbA1c levels (>=5.7%).

Table 4.4 lists the predictive models used in the literature for HbA1c prediction. The studies listed in this table used statistical and mathematical approaches to investigate the correlation between HbA1c levels and clinical variables. However, they did not explore the predictive power of HbA1c levels using deep machine learning techniques using any type of dataset.

Table 4.4: Summary of predictive models used in the literature for HbA1c prediction.

Study	**Method**	**Dataset**	**Year**
Gray et al. (2012)	Logistic regression	Dataset(A1C1)	2012
Handlos et al. (2013)	Multiple logistic regression	Dataset(A1C2)	2013
Stanley Xu et al. (2014)	Linear regression model	Dataset(A1C3)	2014
Kazemi et al. (2014)	Linear regression model	Dataset(A1C4)	2014
S. B. Choi et al. (2014)	SVM and ANN	Dataset(A1C5)	2014
Wells, Lenoir et al. (2018)	Multiple logistic regression	Dataset(A1C6)	2018

4.1.5 Identified Problems Related to HbA1c Prediction in Literature Review

As aforementioned, the related work presented in this thesis depends on reviewing both the datasets and the techniques used by the predictive models employed. To summarise the gaps discussed above, the dataset issues that have been identified in the studies of T2DM diagnosis and HbA1c levels predictions are that they are either of an insufficient size, involve dataset availability, contain imbalanced classes, or are non-diabetes specific. Although machine learning has been intensively used in diagnosing diabetes and discovering its related patterns, it is noticeable that there is a lack of approaches that employ deep learning predictive models using EHR datasets for HbA1c prediction. Furthermore, there is also a lack of studies that have investigated the effects of the temporal nature of EHR datasets and none have used methods to explain the classification decisions made by the predictive models used (blackbox models specifically).

The work described here aims to bridge these areas by applying state-of-the-art predictive models — deep learning specifically — and investigating the temporal effect of time-series data on the performance of the models using a representative and large EHR dataset for HbA1c prediction specifically. Also, this work triggers more investigation into the predictability of current HbA1c levels using more features than those used in the related work studies.

4.2 Mortality Risk Prediction Using Machine Learning

With the help of EHR, clinical data has developed into an interesting frontier for machine learning research. The traditional acuity scoring/measuring methods were accompanied by some conventional machine learning algorithms for the forecast of mortality status. A model by Ghassemi et al. (2015) — that used the Multi-Task Gaussian Process — achieved better results when SAPS scores were added to the input features. Luo et al. (2016) converted MIMIC-II time series data into graphic representations to discover temporal patterns (Saeed et al. 2011). Extracted patterns were then grouped using a non-negative matrix factorisation method. These groups were used with a LR classifier for mortality-risk prediction.

Several studies have measured the mortality risk for a specific health condition such as drug intoxication mortality (Boo and Y. Choi 2020), acute kidney injury (Celi et al. 2012), and liver cancer (Shi et al. 2012). Besides, neural networks have been used for mortality prediction for specific health conditions such as the mortality of pneumonia patients as investigated by Caruana, Baluja and Mitchell (1996) using two different neural network models. Another study by Celi et al. (2012) employed LR, a Bayesian Network (BN), and an Artificial Neural Network (ANN) using the MIMIC dataset to predict the mortality of acute kidney injury patients (1,400 cases) and Subarachnoid Hemorrhage patients (223 cases). They used SAPS and EuroSCORE results to compare with the results of their models. Another study by Shi et al. (2012) compared a neural network model with a LR model for the prediction of mortality following liver cancer surgery with an accuracy of 84%. The conventional neural network models in these studies achieved better performance and outperformed shallow models, such as LR and BN (Clermont et al. 2001).

A study by Silva et al. (2009) proposed a framework to determine the predictors of in-hospital mortality among patients aged 60–104. The data were collected upon admission using questionnaires. A total of 856 patients participated in the questionnaires and provided complete answers. The authors employed the statistical logistic regression method to determine the independent predictors association with in-hospital mortality. The study aimed to find an association between these predictors with the mortality of patients rather than solely predicting mortality.

The predictive models employed in the above studies used traditional approaches and were designed to measure the acuity score of patients with specific health conditions or determine the mortality predictors. However, a recent study by Delahanty, Kaufman and Jones (2018) presented an automated machine learning algorithm, called Risk of Inpatient Death (RIPD), to predict the risk of in-hospital mortality for patients in critical care units. The study used a cohort of 237,173 patient records extracted from EHR data for 131 ICUs across 53 hospitals. The data showed a 9.2% mortality rate for patients admitted at the ICUs for the period 2014–to the end of 2016. The authors used the XGBoost algorithm to estimate the in-hospital mortality risk for adult patients. The XGBoost algorithm was also used by another study by Brajer et al. (2020) on EHR data to predict the mortality risk scores of 75,247 hospitalised patients across three hospitals. Unlike the study by Delahanty et al., the data used in Brajer et al. study reported 2% mortality among the hospitalised patients. The results from both studies are encouraging for the use of machine learning models to predict in-hospital mortality using EHR data.

Several conventional machine learning approaches were employed by Awad et al. (2017) for the early prediction of mortality among ICU patients. The models were trained to predict patient mortality within six hours of admission. A total of 11,722 patient records were extracted from the MIMIC II dataset (Goldberger et al. 2000). The authors investigated the dataset using machine learning, Random Forest, Decision Trees, Naive Bayes and PART, with and without data imputation and oversampling methods. The investigation included employing different input sizes (top 5, 10, 15 and 20 variables). The variables selection was based on the Information Gain (IG) approach. The IG approach measures the variables' contributions to the classes (J. Tang, Alelyani and H. Liu 2014). The RF approach showed that the best results were achieved using vital signs and the top ten variables extracted from the MIMIC II EHR dataset.

A study by Faisal et al. (2020) compared statistical logistic regression models with commonly used machine learning methods, such as RF, SVM and simple neural network, to predict the risk of in-hospital mortality in patients admitted to emergency units. The authors used EHR data from two hospitals (York and Northern Lincolnshire and Goole (NLAG)) with a 4.7% and 6.5% death rate, respectively. As the main objective of the study was to compare logistic regression

with machine learning models, the authors concluded that simple logistic regression models can compete with machine learning models. The logistic regression was shown by the authors to be easier to implement with less effort and complexity.

A deep learning model, the Long Short-Term Memory (LSTM), was used by Harutyunyan et al. (2019) to predict patient mortality using the MIMIC-III dataset (A. E. Johnson et al. 2016). This study used the LSTM for mortality risk classification after 48 hours of patient admission. Their model achieved 86.25% accuracy using the Area Under the Receiver Operator Curve (AUROC) and %51.69 accuracy using the Area Under the Precision-Recall Curve (AUPRC). However, the dataset used in this study was imbalanced (90% discharged home and 10% died).

4.2.1 Discussion of the Datasets and Predictive Methods Used for Mortality Risk Predictions

The work of this thesis focuses on studies that employed machine learning models using EHR datasets and those that were not limited to the prediction of in-hospital mortality risk for patients with a specific health condition. Table 4.5 lists the datasets used in the studies that aimed at forecasting the risk of in-hospital mortality for all patients using EHR datasets.

Table 4.5 also shows the availability of the datasets used in the studies, as well as the the number of attributes and records for each dataset. The mortality rates in the datasets — which reflect the data-balance status and the dates of the datasets — are also presented in the table.

As aforementioned in Chapters 1 and 5, EHR systems were designed to improve healthcare outcomes and were not originally intended for research purposes (Stanley Xu et al. 2014). Patient data stored in EHR systems can be irregular as lab instructions are carried out with different frequencies based on physicians' decisions and patients' visit patterns. It is very common that medical data extracted from EHR systems suffer from problems such as irregularity, incompleteness, and noisy and imbalanced data (Miotto et al. 2016). This implies, the same issues regarding the influence of the datasets that were discussed earlier also apply to these studies.

Table 4.5: Details about the datasets used in the related work for mortality risk prediction.

Dataset	Avail[a]	No. Var[b]	No. Rec[c]	Mortality %	Dataset date
Dataset(MORT1): Delahanty, Kaufman and Jones (2018)	✗	17	237,173	9.2%	2016
Dataset(MORT2): MIMIC II used by Awad et al. (2017)	✓	20	11,722	12.6%	2013
Dataset(MORT3): MIMIC III used by Harutyunyan et al. (2019)	✓	17	42,276	10%	2016
Dataset(MORT4): used by Brajer et al. (2020)	✗	57	75,247	2.7%	2015
Dataset(MORT5) & Dataset(MORT6): York Hospital & NLAG used by Faisal et al. (2020)	✗	9	24,696 & 13,477	4.7% & 5.6%	2014

[a]Avail: Availability. [b]No. Var: Number of variables.[c]No. Rec: Number of records.

Dealing with missing data was one of the major problems discussed in the studies listed in Table 4.2. The authors applied different methods for dealing with this problem. For instance, Awad et al. (2017) employed two techniques for handling missing data as several analyses were conducted in this study using different versions of the datasets. The first technique was filling the missing values with the mean value of each feature. The second technique used was predicting the missing values using the Multiple Imputation algorithm (EMImputation), which uses Expectation Maximisation to statistically find the maximum likelihood of the missing numeric values. Harutyunyan et al. (2019) filled the missing values with the most recent available readings available in the longitudinal data for the patient. However, in the case of complete non-availability, the authors used pre-specified normal readings for each feature.

Table 4.6: Summary of predictive models used in the literature for mortality risk prediction.

Study	Method	Dataset used	Year
Awad et al. (2017)	RF, Decision Trees, Naive Bayes and PART	Dataset(MORT2)	2017
Delahanty, Kaufman and Jones (2018)	XGBoost algorithm	Dataset(MORT1)	2018
Harutyunyan et al. (2019)	Long Short-Term Memory (LSTM)	Dataset(MORT3)	2019
Brajer et al. (2020)	Gradient-boosting, RF and regression models	Dataset(MORT4)	2020
Faisal et al. (2020)	LR, RF, SVM, recursive partitioning, and regression trees (RPART) and ANN	Dataset(MORT5) & Dataset(MORT6)	2020

Other studies attempted to minimise any engineering of features when handling the missing data.

A study by Brajer et al. (2020) did not employ any techniques for missing data imputation. The authors ignored the missing values and replaced them with NA (not available) assuming that the predictive models can learn from the presence and the missingness of values as well of the values of the features. A study by Delahanty, Kaufman and Jones (2018) imputed the missing values with extreme negative values (-9999 specifically). A different approach was used in the study by Faisal et al. (2020). Records with one or more missing values were completely excluded from the development datasets (York and Northern Lincolnshire and Goole datasets), which resulted in a reduction of the dataset size by 24% and 17%, respectively.

Imbalanced datasets was one of the major problems of the datasets used in the above studies. As shown in Table 4.5, all the datasets used in the related studies were imbalanced. The in-hospital mortality incident rate ranged from 2.7% (Brajer et al. 2020) to 12.6% (Awad et al. 2017) in the datasets used in these studies. Hence, the predictive models used in the studies are either trained using the dataset as it is (imbalanced) — such as in (Harutyunyan et al. 2019), (Delahanty, Kaufman and Jones 2018), (Faisal et al. 2020) and (Brajer et al. 2020) — or applied techniques, using over-sampling or under-sampling (Galar et al. 2011), to artificially balance the datasets, such as the study by Awad et al. (2017).

Table 4.6 shows the predictive models used in the in-hospital mortality prediction studies. Most of the employed predictive models were conventional machine learning models. The LR model was investigated by several studies in the literature. In most of these studies, LR models have shown a great capability to predict patient mortality status. In recent years, deep learning models have shown powerful capabilities for analysing and understanding complex clinical data in a variety of medical applications (Goldenberg, Nir and Salcudean 2019). However, deep neural networks models have not been explored much for the prediction of mortality risk for in-hospital patients (Harutyunyan et al. 2019).

4.2.2 Identified Problems Related to Mortality Risk Prediction in Literature Review

In addition to the dataset issues (e.g. missing data and dataset imbalance) that were discussed earlier, the discussions above indicates that there is a real lack of studies that have used EHR datasets for mortality risk calculations, as the studies and the EHR datasets collected are still relatively new as shown in Tables 4.5 and 4.6. Therefore, adopting state-of-the-art machine learning models — and deep learning approaches specifically — to measure in-hospital mortality risk using patient observations routinely stored in EHR and tackling the addressed challenges related to medical datasets, forms one of the main motivations of this work.

4.3 Subsequent Research by Others

There have been several studies that have investigated similar interests during and after performing the work included in this thesis. The studies by Perveen et al. (2019), Islam, Qaraqe and Belhaouari (2020), J.-S. Jang, M.-J. Lee and T.-R. Lee (2019), M. He et al. (2019), and Y. Wang et al. (2020) are studies that have explored the use of machine learning for diabetes or HbA1c prediction. Specifically, the obtained results from the studies by Perveen et al. and Islam et al. reinforce the results obtained in this work. Both studies showed promising results in using machine learning for predicting the diagnosis of diabetes and the levels of HbA1c.

The area of in-hospital mortality risk prediction using machine learning from EHR imbalanced clinical data is also a growing area of research. During and after performing the studies included in this thesis several other studies were conducted, such as the studies by Harutyunyan et al. (2019), Faisal et al. (2020), Brajer et al. (2020) (detailed in the above section). Another study by Y. Gao et al. (2020) who applied machine learning to measure the mortality risk for patients with COVID-19 using the EHR data.

Epilogue

This chapter aimed to present the methods and datasets used in recent studies in the fields of Type-2 Diabetes Mellitus (T2DM), HbA1c elevation levels, and in-hospital mortality risk predictions. As illustrated in this chapter, many studies have investigated the use of predictive models for T2DM prediction. However, few studies have investigated HbA1c elevation levels and mortality risk predictions using EHR datasets. Furthermore, very few studies have investigated the use of deep machine learning approaches as predictive models to overcome these medical challenges. The following chapter will introduce EHR systems and the dataset obtained for the present research project.

Chapter 5

Electronic Health Records and the KAIMRC Dataset

Prologue

Electronic health records (EHR) systems create huge datasets that greatly enrich the field of medical computer science. EHR data can provide an indispensable source of knowledge to help doctors and researchers provide better healthcare services. However, it is challenging to make effective use of data extracted from EHR systems. This chapter aims to introduce and discuss electronic health records (EHR) by providing a definition, a brief history, and outlining the growth in EHR data, before presenting and describing the dataset employed. We also discuss the collection, extraction, preparation, and sampling challenges associated with the dataset used in this thesis.

5.1 Electronic Health Records

According to the International Organisation for Standardisation (ISO), electronic health records (EHR) systems are defined as a repository of patient data stored securely and interchangeably

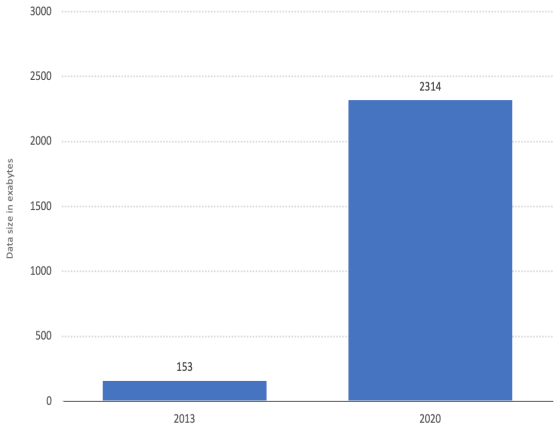
in a digital form (ANSI-ISO 2005; Häyrinen, Saranto and Nykänen 2008). The EHR contains a variety of structured formal medical data such as diagnoses, vital sign records, and laboratory test results, as well as unstructured data such as physicians' notes (Zhao et al. 2017). Daily clinical activities for each patient are stored in EHR systems together with an associated time-stamp to form large sources of longitudinal patient data.

Early efforts in the development of EHR ran alongside the development of computer technology in the 1960s and 1970s (Atherton 2011) and were mainly developed by academic medical centres. However, in its emerging stages, EHR was subject to many operational issues, such as high cost, frequent systems errors, and poor uptake/resistance from physicians (Evans 2016).

In the 1980s, the intended benefits of EHR became more feasible to achieve in the healthcare industry, which encouraged healthcare providers to begin investing in solving incumbent issues to better facilitate the use of the EHR systems. With the advent of more affordable and powerful computer hardware in the 1990s, EHR began a stage of rapid development as the benefits of more accessible medical information became clear to stakeholders. In recent years, EHR has shown an improved capacity for storing, managing and providing access to detailed longitudinal patient data. Crucially, the use of current EHR systems has demonstrated a positive effect on patient care (Manca 2015).

According to a survey by the Ponemon Institute, data held in electronic health records represents 30% of the world's data storage (Z. Liu et al. 2019) and the growth of global healthcare data continues to drastically increase. For example, in 2013, the total size of global EHR was 153 exabytes (1 exabyte = 1 billion gigabytes). This figure is dramatically projected to reach 2,314 exabytes in 2020 (Desjardins 2018) (see Figure 5.1). The astronomical size of current EHR datasets can be exploited to significantly enrich the development of medical analytics, including AI-based analytics. The market size of healthcare analytics in 2019 was USD $14 billion and is expected to reach USD $50.5 billion by 2024 (an annual growth rate of 28.3% (*Healthcare Analytics Market* 2020)).

Figure 5.1: Projected global growth in healthcare data. Adapted from (Desjardins 2018).



Nowadays, data scientists have shown interest in healthcare data especially in the use of electronic health records. Although EHR was initially intended to improve patient care, the datasets generated by EHR systems now represent valuable sources of data for knowledge extraction and decision-support systems. The reuse of EHR data for research purposes is currently considered as a secondary use of health data (Meystre et al. 2017). However, health data analysis has helped speed up scientific medical discoveries and is an important part of clinical research (Botsis et al. 2010). Over the past ten years, EHR data has been increasingly used for a variety of medical analytic tasks and has demonstrated to have a great capacity for supporting medical decision making (Coorevits et al. 2013). However, a notable research gap is the lack of machine learning approaches, especially deep learning, models to assess HbA1c elevation in patients using data extracted from the EHR systems (Harerimana et al. 2019).

5.2 EHR Dataset Challenges

In general, medical datasets present many challenging problems. Although some apply to any medical dataset (e.g. data confidentiality issues); several specific and critical issues are relevant to the use of datasets collected from EHR systems, especially when such datasets are employed

for research purposes, such as data mining research.

EHR medical data is recorded by systems during the clinical treatment and based on physicians' medical decisions. For instance, laboratory tests are ordered by physicians according to the clinical condition of each patient. Therefore, the data extracted from the EHR systems were not originally designed to solve a specific research purpose (Abhyankar, Demner-Fushman and C. J. McDonald 2012). Apart from the standard medical issues, EHR datasets are subject to other issues such as data irregularity, incompleteness, redundancies, and errors. We provide an outline of some of the main challenges related to using EHR data below.

- **Confidentiality**

  Data privacy is one of the most significant problems associated with using medical datasets for research. Health records typically contain confidential and sensitive personal information, such as history of health problems (for example, HIV diagnosis) and personal details (Abouelmehdi, Beni-Hessane and Khaloufi 2018). Naturally, privacy and security regulations are very strict in relation to the use of medical datasets. Maintaining security and privacy in the use of medical datasets is a complex and time consuming. Furthermore, the medical data restrictions and regulations can limit the availability of data used for research purposes.

- **Data Imbalance**

  A major challenge of using clinical data is its imbalanced nature (Batista, Prati and Monard 2004), with most medical datasets usually subject to class imbalance (Galar et al. 2011; L. Zhang, H. Yang and Jiang 2018; Rahman and Davis 2013). To elaborate, these imbalances occur, for example, when the presence of one class of patients in a given dataset outnumbers that of its counterpart class (Longadge and Dongre 2013).

  Several studies highlight that class imbalance is one of the main issues and concerns for applying machine learning algorithms. For example, Weiss and Provost (2001) demonstrated that naturally distributed datasets are suboptimal for use with machine learning

algorithms. To overcome this, the study suggested using different class distributions to those of real world domains. In addition, Mazurowski et al. (2008) found that using imbalanced medical datasets has a detrimental effects on the performance of the neural network model used. A systematic study by Japkowicz and Stephen (2002) concluded that this problem affects the performance of the decision tree algorithms as well as support vector machine (SVM) and neural network algorithms.

In many cases, the minority samples in clinical datasets (usually unhealthy patients with positive results) are more important than those in the majority samples. The distribution of such dataset classes can affect the performance of most classification algorithms. Thus, the predictive models that use imbalanced datasets tend to perform poorly especially when predicting the minority class events (Rahman and Davis 2013).

- **Data Irregularity (Sparsity)**

  In many EHR systems, patient data, such as vital signs and lab tests, is routinely collected and stored with an associated timestamp. The frequency with which these measurements are taken differs among patients, based on the physician's decisions. Patients differ in their visit patterns (e.g., in-patient or emergency visits). It follows that the stay length for each patient varies from a few hours to days, weeks or even months. Moreover, patients suffering from chronic health conditions (such as kidney disease) would have far more comprehensive EHR data associated with them compared to those without chronic illnesses. Therefore, the clinical records for such patients are difficult to identify and extract in a coherent manner (Christensen and Grimsmo 2008).

- **Missing Data (Incompleteness)**

  Missing data poses a major challenge for the use of datasets extracted from EHR systems. The definition of missing clinical data in EHR systems is subject to debate. As mentioned above, data stored in EHR systems are not intended to solve a specific research problem (Pivovarov 2015; Wells, Chagin et al. 2013). Hence, the absence of some variables in EHR systems is not at random. Simply, the absence of the variables in the EHR system is mainly reliant on clinical protocol and decisions. EHR datasets tend to be subject to missing data

due to the inherent irregularities in making observations of a patient's condition. However, from the data science prospective, the lack of data entries for any variables used in a medical predictive model is considered as being missing.

- **Data Entry Errors**

  While computerised medical systems have shown great success in improving human health, the issue of clinicians erroneously entering data onto EHR systems poses a major problem. To overcome this, various alert mechanisms have been formulated to prevent data-entry errors on EHR systems by detecting common data-entry mistakes made by clinicians. However, the reliability and efficiency of these techniques used in EHR have been argued as being inadequate and insufficient (Hecht 2019).

- **Data Heterogeneity**

  As mentioned above, the clinical data stored on EHR systems contains both structured and unstructured data types (Pivovarov 2015). For instance, laboratory tests and vital sign readings are stored in structured forms while clinicians' notes about their patient observations are stored in an unstructured form as text or audio, and clinical imaging results are recorded on EHR systems as images or videos.

- **Data Standardisation**

  Hospitals around the world use different standards for disease classifications and coding, the naming of laboratory tests, and units. Furthermore, there are different versions within the same standard. The latest version of the International Statistical Classification of Diseases and Related Health Problems (ICD), provided by the World Health Organisation (WHO) is the tenth version, however, a newer version has been adopted by the WHO and is planned to be take effect by 1 January 2022 (World Health Organisation 2020). As an example, the blood glucose levels are measured in $mmol/L$ in some hospitals whereas in other hospitals it is measured in mg/dL (more details about this issue will discussed in Chapter 7).

5.3 King Abdullah International Medical Research Center Dataset

5.3.1 Dataset Population

The King Abdullah International Medical Research Center (KAIMRC) dataset was collected by the Ministry of National Guard Health Affairs (NGHA). KAIMRC is one of the Middle East's leading institutions in health research. The data has been extracted from the main National Guard Hospitals located in the three most populated regions and is sorted by population, western, central, and eastern regions. The latest statistics from the Saudi Central Department of Statistics and Information (CDSI) shows that 66.6% of the population in Saudi Arabia live within these three regions (Central Department of Statistics & Information (CDSI) 2018).

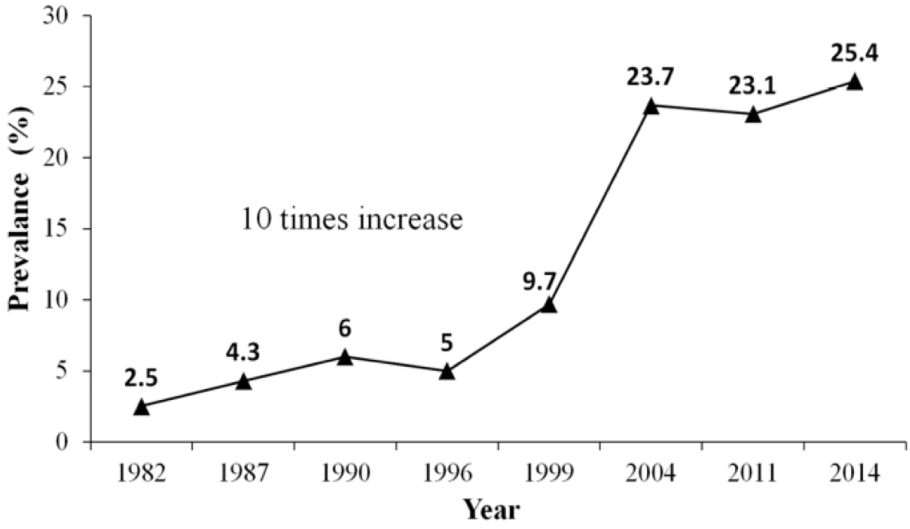
The prevalence of diabetes in Saudi Arabia is rapidly increasing and is recognised as one of the main causes of death in the country (Abdulaziz Al Dawish et al. 2016). Figure 5.2 shows the increase of the prevalence of diabetes in Saudi Arabia between the 1980's to 2014. Furthermore, in 2016, Saudi Arabia was ranked by the World Health Organisation (WHO) as having the second-highest prevalence of diabetes in the Middle East and seventh highest in the world, with a diabetes prevalence rate of 18.3%. Unsurprisingly, estimates show that healthcare expenditure on diabetes treatment in Saudi Arabia has risen 500% since 1998.

5.3.2 KAIMRC Dataset Collection

The KAIMRC dataset used in this work was originally extracted in two parts based on approved KAIMRC projects. The first part provides patient data from 2010 to mid-2015 for the following projects:

- Project 1: Diabetes Early Warning System, Research Protocol SP14/042.
- Project 2: Finding the Common Related Diseases With Diabetes Using Data Mining Association Techniques, Research Protocol SP15/064.

Figure 5.2: Diabetes prevalence in Saudi Arabia from 1982 to 2014. Adapted from (Abdulaziz Al Dawish et al. 2016).



These projects were extended under project extension number RYD-17-417780-187503 to collect more data. Thus, the second part of the collected dataset provides patient data from 2016–to the end of 2018. The first part of the dataset was obtained in mid-2017 and the second part in mid-2019. The data has been provided to be used in this thesis under the following conditions: (i) the provided data should be used for research purposes only; (ii) it should not be used in other projects without permission.

The National Guard hospitals use the tenth version of the ICD as the standard classification of diagnosis, symptoms, clinical signs and laboratory orders (World Health Organisation 2004).

5.3.3 Profile for KAIMRC Dataset

This section provides a general overview of the KAIMRC dataset and its statistical characteristics. In total, the collected dataset (parts 1 and 2) consists of 122,327 unique patients who made 765,318 visits from 2010–2018. During this period, more than 87 million clinical readings

(excluding vital signs readings) were taken and stored in the EHR systems[1]. The profile of the collected dataset is shown in Table 5.1.

Table 5.1: Profile for KAIMRC datasets (part 1 and 2).

Characteristics	**Overall dataset**
Total number of unique patients	122,327
Total number of visits	765,318
Number of features	Over 500
Total number of clinical readings	> 87 million (excluding vital signs)
Overall period	2010 to end of 2018

Table 5.2 shows more details about the two parts of the KAIMRC dataset. The first part of the dataset contains data for 12,499 unique patients with 14,609 visits. More than 41 million clinical readings were recorded during these visits. The second part contains data for 114,057 unique patients who made a total of 750,716 visits. Overall, the dataset provides details for over 46 million observed clinical readings for more than 500 unique laboratory tests.

Table 5.2: Profile for KAIMRC datasets' parts.

Dataset	**Number of unique patients**	**Number of visits**	**Number of clinical readings (in millions)**	**Period**	**Obtained year**
Part 1	12,499	14,609	> 41 excluding vital sing readings	2010 - mid 2015	2017
Part 2	114,057	750,709	> 46 excluding vital sing readings	2016 - end of 2018	2019

With regards to diabetes prevalence in the collected dataset, Figures 5.3 and 5.4 illustrate the distribution of patients with diabetes in both parts 1 and 2 of the KAIMRC dataset. This includes patients with one or more type of diabetes. Since parts 1 and 2 of the dataset are used independently in the studies presented in this thesis, we present them in separated figures.

[1] Excluding the period from mid-2015–the end of 2015.

Figure 5.3: Diabetes diagnosis distribution in KAIMRC dataset part 1.

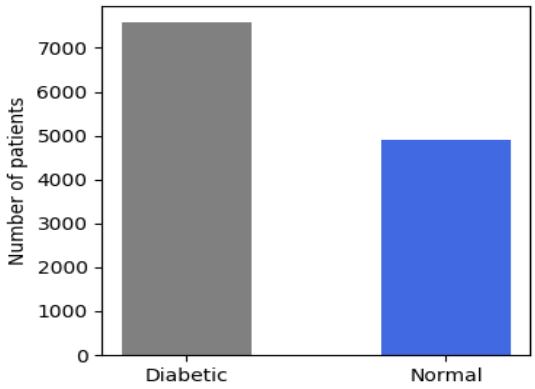


Figure 5.4: Diabetes diagnosis distribution in KAIMRC dataset part 2.

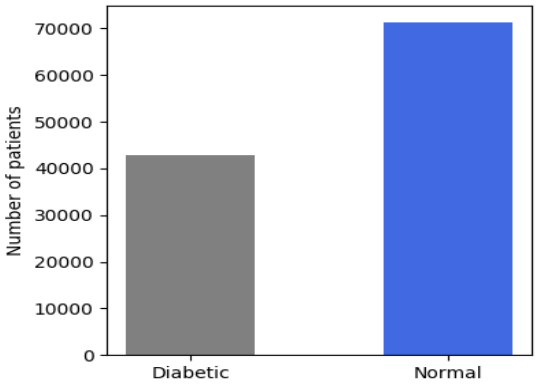
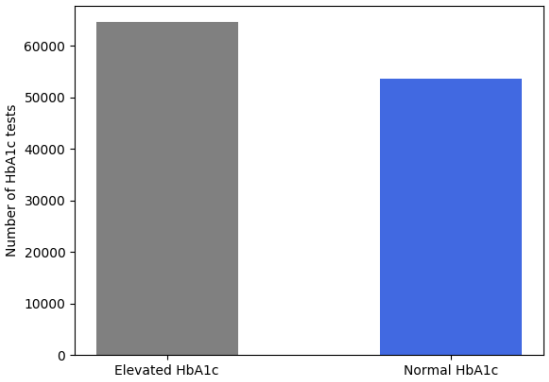


Figure 5.3 shows that 60.7% (7,587/12,499) of the patients in part 1 of the dataset had diabetes while 39.3% (4,912/12,499) did not have it. For part 2 of the dataset, 37.5% (42,729/114,057) of patients had diabetes while 62.5% (71,328/114,057) did not have any type of diabetes as shown in Figure 5.4.

With regards to elevated HbA1c levels (>= 5.7%) among the patients without hyperglycemia, visits by patients diagnosed with any type of diabetes (42,729 patients) were excluded. Since only part 2 of the dataset is used for the investigation of elevated HbA1c prediction in the non-diabetic population, Figure 5.5 shows the distribution of the HbA1c test results in this part (for non-diabetic patients). For patients without diabetes (71,328), 5,147 of these patients had not undergone any HbA1c testing. Therefore, HbA1c levels were tested in 66,181 unique patients that were not having diabetes diagnosis in the EHR systems. Finally, 118,156 unique HbA1c tests were carried out on patients with an average HbA1c test count of 1.8.

The dataset also provides information about the patient gender distribution of the KAIMRC dataset. Figures 5.6 and 5.7 illustrate the patients' gender distribution in the KAIMR dataset

Figure 5.5: HbA1c levels distribution in KAIMRC dataset part 2 for patient's visits (without hyperglycemia diagnosis).



parts 1 and 2. Specifically, there were slightly more female patients than male patients in both parts of the dataset (55.83% vs 44.17% for part 1 and 55.45% vs 45.55% for part 2, respectively)[2].

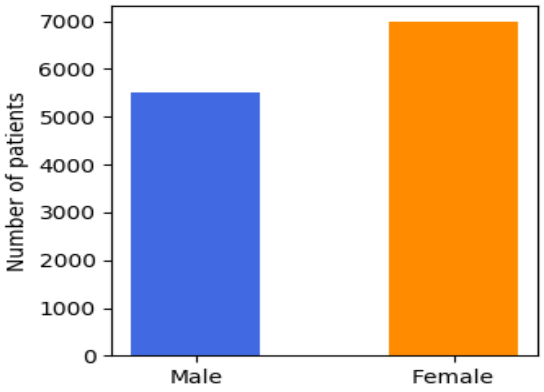


Figure 5.6: Gender distribution for patients in KAIMRC dataset part 1.

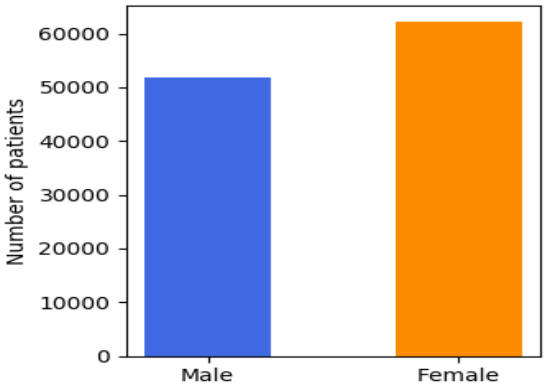


Figure 5.7: Gender distribution for patients in KAIMRC dataset part 2.

[2] A few patients are reported having an undefined or missing value for the gender

Figures 5.8 and 5.9 illustrate the number of visits made over patients' gender in the KAIMR dataset parts 1 and 2. Unsurprisingly, female patients made a higher number of visits overall than male patients in both dataset: 56.18% vs 43.82% in part 1 and 56.33% vs 43.67% in part 2, respectively.

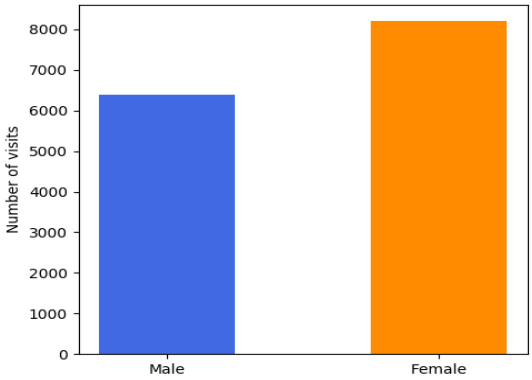


Figure 5.8: Number of visits made by the patients in KAIMRC dataset part 1.

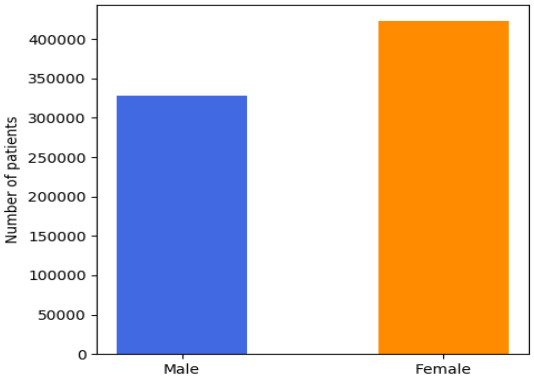


Figure 5.9: Number of visits made by the patients in KAIMRC dataset part 2.

As a preliminary work, we applied several machine learning algorithms to predict the clinical diagnosis of patients with Type-2 Diabetes Mellitus (T2DM) to validate the KAIMRC dataset. The models employed used 30 features/variables that were available in the KAIMRC dataset part 1. The models employed achieved promising results using data of patient visit. However, this thesis focuses on investigating the prediction of HbA1c levels employing advanced machine learning approaches using EHR (KAIMRC) dataset to help identify those patients at risk of diabetes (and pre-diabetes).

5.3.4 Main Characteristics of KAIMRC EHR Dataset

The KAIMRC EHR dataset is a large unique dataset that contains details about patient visits for the three main National Guard hospitals. Most importantly, it contains detailed clinical

diabetes diagnoses, rich patient background information, and full patient history with time stamps detailing all patient visits. In this work, the KAIMRC dataset is provided anonymised (without the details that can be used for identifying the patients).

In addition, the KAIMRC dataset contains the personal details of every patient, such as age and gender, along with vital signs data and lab test results for every visit. As patient clinical data is usually processed at irregular times before being stored in the hospital record systems, lab test results such as Blood Urea Nitrogin (BUN), Total Cholesterol (CHOL), and Mean Corpuscular Haemoglobin (MCH) are collected along with their timestamp values. The dataset also holds time-stamped values for other vital signs such as Body Mass Index (BMI) and hypertension readings.

One of the main characteristics of the KAIMRC dataset is its size. While it is arguable what is considered optimal size of datasets being used in predictive models, comparing to the studies reviewed in Chapter 4, the KAIMRC dataset is larger than any other datasets used for HbA1c levels prediction (refer to Table 4.4). The availability of such large dataset makes it possible to train advance machine learning techniques, e.g. deep learning models.

Data balance is another important characteristic of the KAIMRC EHR dataset. Although most medical datasets are usually found to be imbalanced, surprisingly, comparing to the studies presented in 4, the KAIMRC dataset shows better distribution (balance) for the classes with regards to HbA1c levels (this issue will be discussed in detail where applicable for each population used in the studies included in this work).

5.3.5 Main Challenges Presented by the KAIMRC Dataset

As previously outlined, one of the main challenges posed by the KAIMRC EHR dataset is its availability. Although it is not publicly available, researchers can request access to the dataset by making an official access request to the National Guard Health Affairs (NGHA) in Saudi Arabia. It is important to highlight that the analyses performed in this work employ structured

numerical or categorical data only. Non-numerical variables that have a fixed set of possible values were encoded for categorical data such as patient gender and blood type and antibody tests.

As mentioned above, the KAIMRC dataset shows better balance when compared to the studies reviewed in Chapter 4 for the classes distribution with regards HbA1c levels. However, the dataset does suffer from the data imbalance with regards to mortality rates as well as other common EHR problems such as incomplete data, data sparsity, and data errors, as we examine next.

Incomplete data poses a major challenge in using the KAIMRC dataset. As noted earlier, missing data in datasets collected from EHR systems is not due to input errors or extraction issues. In fact most of the apparently unavailable medical data is missing because these data points do not form part of the clinical practice or protocol followed by a specific physician for the treatment or diagnosis of a specific patient suffering from a specific symptom(s). At the same time, common missing data interpretation approaches are not always justifiable from clinical perspectives. However, a unique approach to overcoming the challenge of missing data (missing data interpretation) will be discussed in detail in the analyses associated with each study presented in this thesis.

Further, data sparsity is an issue in the KAIMRC EHR dataset. Specifically, while the dataset contains hundreds of data points relating to patient visit history and thousands of vital-sign and lab test readings, it also features as few as a single patient visit, a single vital sign data point, or a single lab test reading. It is also important to highlight that the dataset also contains data errors such as bad characters, use of different scales, and extreme values, which present an additional challenge for the data preparation stage.

Although KAIMRC is a unique dataset, using it for the present project introduces new challenges. Specifically, as the dataset was collected for two separate stages, there are two main differences between the two parts of the dataset. First, as mentioned in the Scope section in the Introduction

chapter, one of the main differences is the absence of particular variables in one or other of the two parts. For instance, while systolic and diastolic blood pressure readings are available in the first part, they are missing in the second.

Second, another important difference between the two parts of the dataset concerns the sources of each hospital's data. Specifically, the first part of the dataset was sourced from hospitals across all three regions (western, central, and eastern regions), whereas the data in the second part of the dataset was sourced from only two regions (western and central). The main differences between the two parts of the dataset are shown in Table 5.3. Besides these differences, data from mid-2015–end of 2015 is/was not available.

Table 5.3: Main differences between for KAIMRC datasets' parts.

Difference	**Dataset part 1**	**Dataset part 2**
Patient vital signs data	BMI and systolic and diastolic blood pressure	Only weight and height are available
Clinical diagnosis	Only diabetes	Variety of clinical diagnosis (including diabetes)
Patient discharge details	✓	✗
Patient visit type	✓	✗
Medication details	✗	✓

5.3.6 KAIMRC EHR Dataset Pre-processing

The KAIMRC datasets (parts 1 and 2) are extracted from the EHR systems in CSV file format. The data is/was organised in a semi-structured format that include many redundancies, errors, and missing values. For part 1, the CSV files contain all patient information (personal details, diagnosis, vital signs, and lab tests). Tables A.1 and A.2 in section A.1 in Appendix A show the structure of the KAIMRC data (part 1) as originally collected. For part 2, the CSV files contain patients' personal details, diagnosis, vital signs, and lab tests. The files are‘ linked by patient medical records number (MRN) and visit ID. See Tables A.3, A.4, A.5 and A.6 in section A.2 of Appendix A to view the structure of the KAIMRC data files for part 2 as originally collected.

For each lab test result recorded at a patient visit, details such as patient number and admission details are redundantly recorded. This redundancy is ascribed to the nature of the data collected

from clinical EHR systems as the frequency and order of the clinical procedures vary from one patient to another. The same vital sign reading, or lab test can be ordered more than once during a particular patient visit. In clinical practice, it is very common that some lab tests and vital signs are recorded frequently, as physicians may have ordered them in this manner to collect longitudinal vital sign data. Besides, it is very common that some measures are not ordered at all during the entire duration of a patient's treatment, hence the missing values.

5.3.7 Features Selection and Preparation

The KAIMRC dataset provides over 500 variables (features) available for extraction as well as personal details on patients and admissions. A preliminary features selection (filtering) approach was used to initially prepare the data for further features selection methodologies that we apply in each study to achieve the objectives of this thesis.

To achieve initial filtering, the most commonly available laboratory tests in both parts of the dataset were selected. Tables B.1, B.2, and B.3 in Appendix B show the top laboratory tests available in the dataset. The most available laboratory tests were calculated based on the number of unique patients who have undergone these laboratory tests at least once. A description of the laboratory tests and their clinical codes are also provided in Appendix B. The clinical codes for the laboratory tests are presented according to the 10th version of the ICD.

In addition to these features, some missing features can be calculated from other available features. For instance, two features important for this study: BMI (for part 2 of the dataset) and Non-High Density Lipoprotien (Non-HDL) are not available. However, these features are added after being calculated using the appropriate formulae.

The selected features represent a mixture of numerical data (e.g. patient age) and non-numerical data (e.g. gender and diagnosis). Because the applied machine learning models in this thesis accept only numerical values, the categorical features that have definite possible non-numerical values were converted into numerical values using the appropriate encoding techniques (details will be provided where applicable for each study in the following chapters).

5.3.8 Interpretations for the Issues in KAIMRC Dataset

As discussed above, in clinical practice, clinical data, vital signs readings, and lab test orders specifically, are taken based on patients' condition and physicians' decisions. Therefore, a large number of readings are not taken (missing values), though this is standard for EHR datasets. Moreover, some of the clinical measures were collected at different frequencies. And not all clinical features are sampled at the same intervals, some clinical measures are taken on an hourly, daily, or even monthly basis.

After discussions with clinicians and laboratory specialists and reviewing the related studies, we adopted the following approach to overcome the missing data problem. In case of missing values for features/variables related to first-day patient visit data, we consider the first available value in that visit. If no values are available for an entire visit, the first next available value in the following visits closest to the missing value is considered (prior values). This approach was also applied by Wells, Lenoir et al. (2018). However, in case the value of interest is still missing in subsequent patient visits or there are no subsequent visits made by the patient, the last available value from the previous visits is considered. Otherwise, the value will be considered missing and a null value used.

To interpret the issue of data sparsity, we apply the following approach. When there are multiple values for the same variable in the first day, the average value of that variable is calculated and considered. For the data recorded in the EHR systems for patient visits with more than one day, the first day data is considered.

With regards to the presence of erroneous values in patient records (extreme values and the fields that contain unexpected entries), the offending value(s) are removed and considered as missing (null value). Few variables (in KAIMRC dataset part 1 specifically) were found to have high number of erroneous values (i.e. Red Blood Cell (RBC), White Blood Cell Count (WBC), and Bilirubin laboratory tests). In this case, we used the clinical assessments (normal, abnormal, abnormal low, and abnormal high) for those variables.

Another problem is the use of different scales by laboratory specialists for recording the readings for some tests. For instance, the Mean Cell Haemoglobin Concentration (MCHC), Haematocrit (Hct) and Haemoglobin (Hgb) blood tests are stored in the KAIMRC EHR (part 2 specifically) systems as percentage or decimal. To interpret this issue, we first discussed these issues with the clinicians and then unified the scales (as decimals) for the values of those variables.

5.3.9 Sampling Approaches

As this thesis consists of several studies, we required different sampling approaches when processing the KAIMRC dataset to achieve the research objectives. We use different sets of predictors throughout this thesis according to the their availability in KAIMRC EHR systems and the methodological approach used in each study.

This produced five different data subsets (A, B, C, D, and E). Table 5.4 illustrates the general profile of these experimental data subsets and highlights where these data subsets are used in this thesis. Subsets A and E were sampled using KAIMRC dataset part 1 while subsets B, C, and D were sampled using part 2. More details about these data subsets will be presented at relevant points in this thesis, chapters 6, 7 and 8.

5.3.10 Contribution to the Creation of the KAIMRC Dataset

Besides seeking for the proper approvals for the KAIMRC dataset access, the contributions to the creation of the dataset used to achieve the research objectives of this thesis include:

- Data preparation including extraction and transformation (time-series data specifically) into a structured format to be used as input for the machine learning models used.
- Reforming the irregular EHR data into a unique sequential input structure to help add behavioural information that will improve the model's predictive performance.

Table 5.4: Profile for the experimental data subsets used on this thesis.

Data subset	**Sampled from**	**Size**	**Number of features included**	**Temporal details**	**Chapter / Section used in**
Data subset(A)	KAIMRC part 1	13,317	78	✗	Chapter 6 section 6.1
Data subset(B)	KAIMRC part 2	36,378	6	✗	Chapter 7 section 7.2
Data subset(C)	KAIMRC part 2	18,844	6	✓	Chapter 7 section 7.3
Data subset(D)	KAIMRC part 2	18,844	27	✓	Chapter 7 section **??**
Data subset(E)	KAIMRC part 1	3,557	86	✗	Chapter 8

- Since the EHR data is not specifically designed for use in addressing specific research problems, associated issues such as erroneous and missing data represent major difficulties for data science and especially in the medical domain. To overcome this, data cleansing is performed on the KAIMRC dataset (this includes fixing the erroneous values, duplicate records and missing data).

Epilogue

This chapter has provided a brief introduction to EHR data and outlined the main challenges of using such datasets. Next, we outlined the nature of the KAIMRC dataset in terms of its population makeup, overall characteristics, data-collection issues, and pre-processing tasks. Then, it explained our features selection approach and the different subsections of the KAIMRC dataset. Finally, we detailed the sampling approaches employed. The following chapter explains how this work bridges the research gaps identified (related to predictive models and datasets used) in the literature by employing advanced machine learning approaches to process the dataset presented in this chapter for identifying patients with risk of diabetes via Glycated Haemoglobin (HbA1c) prediction.

Chapter 6

Collaborative Denoising Autoencoder for Diabetes Risk Identification via Glycated Haemoglobin Prediction

Prologue

This chapter provides an investigation about using state-of-the-art predictive models for predicting the diabetes levels of HbA1c to paves the way for performing detection and monitoring of Type-2 Diabetes Mellitus (T2DM). KAIMRC dataset part 1 is used in this chapter.

This chapter sought to address the research question below (mentioned in the Introduction chapter):

- Can machine learning models assist in predicting the levels of Glycated Haemoglobin (HbA1c) using typical EHR data for patients (diabetic and non-diabetic)?

Declaration: Description of this study as presented in this chapter are largely as published on the following publication: Alhassan Z, Budgen D, Alessa A, Alshammari R, Daghstani T, Al Moubayed N. *Collaborative Denoising Autoencoder for High Glycated Haemoglobin Prediction.* In International Conference on Artificial Neural Networks 2019 Sep 17 (pp. 338-350). Springer, Cham (Alhassan, Budgen, Alessa et al. 2019). The references and notations have been altered, cross-references have been added and some stylistic changes have been made for the consistency throughout this thesis.

6.1 Introduction

As aforementioned in Chapter 2, the level of HbA1c is strongly related to the average glucose concentration in the blood and the life span of the red blood cells. The normal red blood cells of the human body can last for two to three months before being reproduced. Hence the level of HbA1c can indicate the average level of blood glucose over the whole period of the life span of the red blood cells (Larsen, Hørder and Mogensen 1990; A. D. Pradhan et al. 2007). This can provide physicians with an important long-term measure for blood glucose levels (Ackermann et al. 2011).

In this chapter, we introduce the first study that employs deep learning models to predict the level of HbA1c using patient's data stored in the EHR systems. The study uses routinely collected data from hospital patients to predict the level of HbA1c. We introduce a novel deep learning based framework to predict the level of HbA1c using a significantly large and unique clinical dataset (KAIMRC dataset). According to the IEC classification of the test level, we formulate the risk of the HbA1c test classification problem as a binary classification problem (patients with less than a 6.5% HbA1c level being coded as low HbA1c, and for 6.5% or more being coded as high HbA1c). The main contributions of this work presented in this chapter are:

- Introduces a novel collaborative autoencoder framework for the HbA1c normal and abnormal levels prediction.
- Investigates the use of the routinely collected clinical patient data from EHR systems as predictors for HbA1c levels.

6.2 Method

In this work, we investigate employing several machine learning models to predict the level of HbA1c using patient's EHR data. We compare our results against popular base-line models: Support Vector Machines (SVM) and Logistic Regression (LR). We also compare our results to deep learning approaches such as Multi-Layer Perceptron (MLP) and Denoising Autoencoder (DAE) (LeCun, Bengio and Hinton 2015; Hochreiter and Schmidhuber 1997). These models are trained and evaluated using KAIMRC dataset part 1 (detailed in Chapter 2). This work reports an investigation into the performance of machine learning models to predict current HbA1c levels as a binary classification problem (patients with low levels of HbA1c are coded with zero and those with high levels are coded with one).

Autoencoder (AE), detailed in Chapter 3, had several successes in diverse areas of applications, and especially recently with the development of deep variations (Al Moubayed et al. 2016; S. Gao et al. 2015). In the medical field, autoencoders were mainly used to analyse medical imaging data including: removing the noise (Gondara 2016), data analytics (Shin et al. 2011), and outlier detection (Baur et al. 2018). In this work, we investigate a unique collaborative autoencoder framework for high HbA1c levels prediction.

To increase the separability between the classes, high vs low HbA1c levels, the framework generates new features from each class separately by modelling directly the data that belongs to a given class. This is motivated by the success of pre-training in deep learning models (P. Vincent et al. 2010; Goodfellow et al. 2016), however we use a separate model per class to reduce the within-class noise and increase between-class separability. These two models (collaborate) by combining

their outputs together to form the input to a third classification model. The details about the collaborative autoencoder model is presented in Model and Experimental Setup subsection. The next section details the profile for KAIMRC dataset, features selection and input preparation.

6.2.1 Dataset Profile, Features Selection and Preparation

Each patient visit is described by a set of measures. These measures are represented as episodes. Episodes contain the data of irregularly collected vital signs and lab readings. In addition to this, the patient details and visit details (e.g., gender, age, visit type and service provided) are integrated into the episodes. For a patient with an in-patients visit type, only the data for the first day was considered. Cases with values of less than 0.1 of HbA1c are considered to be erroneous readings and have been excluded. This resulted in reducing the dataset, KAIMRC data (subset(A)), size from 14,609 down to 13,317 cases (Table 6.1). Refer to Appendix B for more details (ICD10 code and description) about the variables employed in this data subset.

Table 6.1: Statistics of KAIMRC data subset(A)

Characteristic	**Overall**
Number of patient visits	13,317
Number of features	78
Number of different health conditions	99
Number of patient visit types	4
Number of discharge types	8

We use 78 features for our analysis: gender, age, service, specialty, visit type and 73 vital signs and lab results that are available in the KAIMRC dataset part 1. Some of these features are collected frequently on an hourly basis, such as vital signs. In these cases, the average value for the readings on that day is used instead. 58% of the KAIMRC dataset is labelled as high level: patients with a 6.5% HbA1c level or more, while the remaining 42% are labelled as low level (less than a 6.5%).

An integer encoding method was used to encode the values of categorical features such as age and gender. Data standardisation is used to change the distribution of the features' values so that they are centred on 0 and a standard deviation of 1.

We measured the correlation between the HbA1c level and the 78 features using Pearson Correlation Coefficient (PCC) and Spearman approaches (Benesty et al. 2009). The result shows positive linear correlation for 44 features and zero or negative correlation for the remaining 34 features. There are three features, age, Triglycerides (Trig) and Random Blood Sugar (Glur), also referred as RBS, with highest correlation absolute values. Figure 6.1 shows the class distribution with regards to these three features.

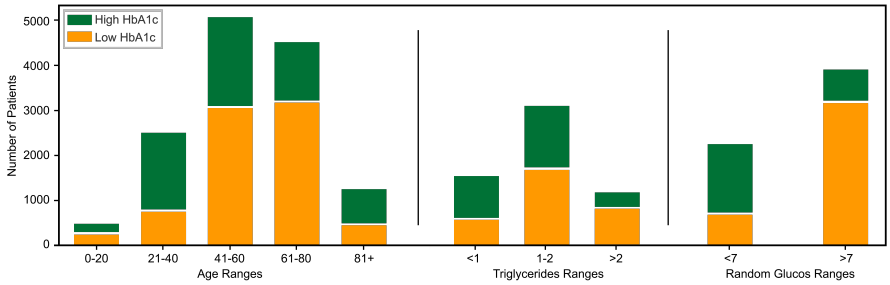


Figure 6.1: Classes distribution over patients age, random glucose and triglycerides.

In general, mining clinical data is made difficult by several problems such as missing data, variety of lengths, and irregularity. For instance, the percentage of missing values for Triglycerides and Random Glucose is 54% and 51% respectively. As unobserved clinical readings (missing values) are never known precisely (Cro et al. 2020), we avoided using techniques to interpolate this problem.

Predicting the HbA1c level using only general clinical data is very challenging. There are many factors that affect HbA1c level and stability such as improved diet and physical exercises. Changes in patient lifestyle, structured exercise training specifically, is known to have a significant effect on the level of HbA1c (Sanghani et al. 2013). Furthermore, the significant amount of missing data in most of the medical datasets forms another major challenge. Figure 6.2

demonstrates the challenge of separating the data between the two classes: high and low level HbA1c, by visualising a two dimensional projection of the data using t-SNE (Maaten and Hinton 2008).

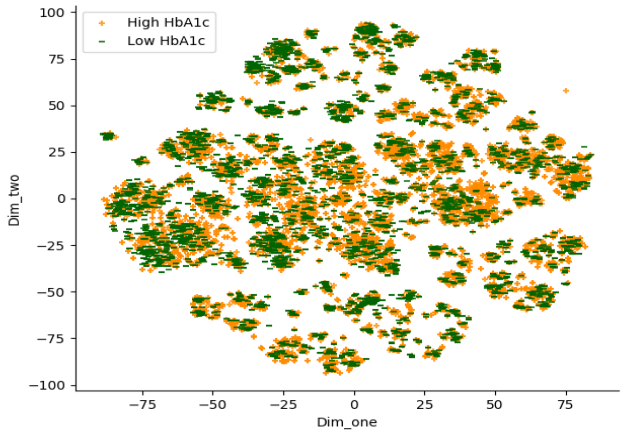


Figure 6.2: Projection of the row data onto two dimensional space using t-SNE.

6.2.2 Model and Experimental Setup

Figure 6.3 demonstrates the collaborative denoising autoencoders (Col-DAE) framework piloted here. Autoencoder1 models the low HbA1c level class, while Autoencoder2 models the high HbA1c level class. The features of latent space of both models are then merged and fed into the MLP classification model. The MLP model is trained to predict the level of HbA1c. To classify a sample, we feed it to both pre-trained autoencoder models with their outputs merged and fed to the MLP model for prediction.

Our framework (Col-DAE) consists of two denoising autoencoders and one Multi-Layer Perceptron (MLP) model. The first DAE (Autoencoder1) models the low HbA1c level while the second DAE (Autoencoder2) models the high HbA1c level. Each DAE model has three hidden layers for the encoder, as shown in Figure 6.3. Prior to the encoder, an isotropic Gaussian distributed noise is added to the input layer (P. Vincent et al. 2010). The number of neurons

for the layers in the encoder are 90 , 120 and 130 respectively. Each DAE model consists of two hidden layers for the decoder. The first decoding layer (90 neurons) takes the embeddings in the latent space as an input while the output layer has the same size as the input. Tanh activation function is used in all encoder and decoder layers.

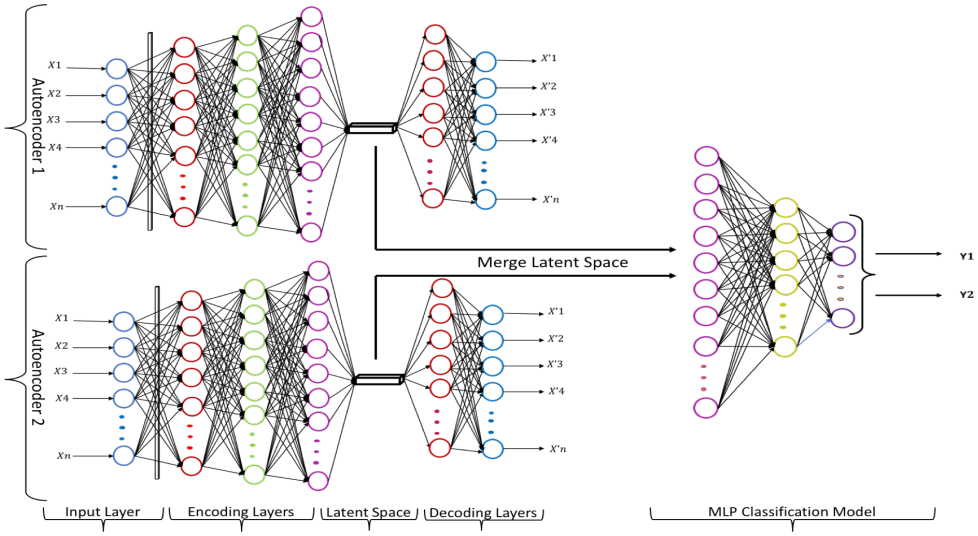


Figure 6.3: Used collaborative-denoising autoencoders (Col-DAE) framework.

An MLP model is used as the classification model. It consists of three dense layers. The first merges the latent spaces from the two DAEs with 260 neurons (130 from each DAE). The second and third layers of the MLP classification model has 70 and 32 neurons respectifly. Relu activation function is used in both layers and Sigmoid for the output layer used for the two classes prediction.

The DAEs (Autoencoder 1 and Autoencoder 2 in Figure 6.3) are pre-trained and validated to reconstruct the input using 80% and 10% of the data respectively. The remaining 10% of the data is kept as the test set. The MLP classification model is trained and validated using the training and validation sets along with the associated HbA1c labelled levels. Each test sample is fed to both DAEs and the embedding outputs from both DAEs are merged to form the input

vector of the MLP classification model. These tasks (pre-train the DAEs, train the MLP and testing the Col-DAE model) are repeated 10 times using a 10-folds cross-validation approach to ensure that all data points are included in testing the model. The DAEs of the Col-DAE model are trained for 100 epochs using an Adam optimiser with Mean Squared Error as the loss function. The MLP of the Col-DAE model uses same optimiser, loss function and number of epochs for training.

The optimisation for the machine learning models (conventional and deep) used in this thesis was performed by analysis of the empirical results obtained using the KAIMRC dataset. For instance, the optimisation process for the deep learning models (Col-DAE, MLP, and DAE used in this chapter) involved tuning the neural network structure (e.g. the number of hidden layers and neurons) and hyperparameters such as the activation functions (detailed in chapter 3), optimisers, and loss functions.

6.3 Results

We explored all models with different feature sizes (using the top three correlated features and the 78 originally collected). Because deep learning approaches work with high dimensional data, the MLP, DAE and Col-DAE models were not investigated using three features. For the purpose of fair comparison, these models were employed using the same data pre-processing, training and testing techniques used for the Col-DAE model.

In addition to the accuracy, we report F1, F1 Weighted, Recall and Precision measures to evaluate the performance of the used models. The Col-DAE model was investigated using different combination of regularisers, activation functions, dropout rates, learning rates, and optimisers. We only report the results with the best performance as per the reported measures.

The performance metrics for predicting the level of HbA1c, obtained using the compared models: SVM, LR, MLP, DAE and Col-DAE with different feature sizes, are presented in Table 6.2. All models show better performance when using the 78 features despite the linear correlation

values for the features except for DAE models. We report an F1-score of 73.34% and 65.63% F1-Weighted measures for DAE. However, the DAE models show clear signs of over-fitting. The large difference between the two measures is explained by the 79.08% recall and 68.48% precision. The SVM model using 78 features showed slightly higher recall accuracy (83.22%) than the Col-DAE and MLP models. However, the same SVM model showed lower precision accuracy (74.94%). This large difference between the recall and precision measures (high recall and low precision) indicates that the SVM and DAE models using 78 features showed biased classification behaviour toward the positive samples.

The rest of the models (SVM, LR, MLP and Col-DAE) do not show any bias towards any of the classes. The SVM and MLP models with 78 features achieved competitive performance with an F1-score of 78.84% and 79.72%. However, the SVM model achieved promising performance, using three features only, with 76% F1-score.

Table 6.2: Performance of classifiers for HbA1c risk prediction

Model	**Features Size**	**F1-Score**	**Accuracy**	**F1-Weighted**	**Recall**	**Precision**
SVM	3	0.7609	0.7172	0.7162	0.7724	0.7499
	78	0.7884	0.7398	0.7356	0.8322	0.7492
LR	3	0.7168	0.6526	0.6476	0.7543	0.6830
	78	0.7574	0.7113	0.7098	0.7733	0.7424
MLP	78	0.7972	0.7588	0.7573	0.8146	0.7827
DAE	78	0.7301	0.6563	0.6445	0.7985	0.6737
Col-DAE	78	**0.8109**	**0.7760**	**0.7751**	0.8240	0.7987

Table 6.2 shows that the Col-DAE model using 78 features achieved better results than the compared base-line models, with 81.09% F1-score and 77.60% of accuracy. Figure 6.4 summarises the 10-folds results of the reported measures achieved by the used models. Figure 6.4 shows small variation between the folds and especially in F1-score which demonstrates the consistency of the Col-DAE's performance. The differences between the obtained F1-scores for Col-DAE with 78 features and the comparative models are statistically significant (P-value (using paired t-test) is 0.0007 with SVM, <0.005 with LR and DAE and 0.028 with MLP).

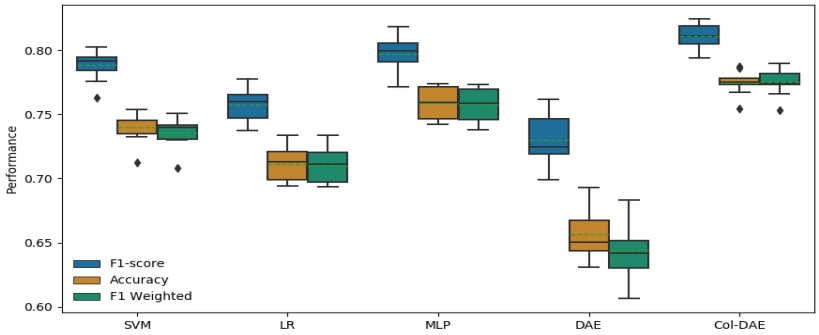


Figure 6.4: Box plot of the detailed performance for the models using all features.

6.4 Discussion and Conclusion

Our framework is trained using patient clinical data from patients visiting the hospitals for a variety of health conditions. Despite the large number of missing values, the SVM achieved 76% F1-score using three features. The Col-DAE outperformed the base-line classifiers and achieved 81% for F1-score using 78 features. Due to the lack of similar studies and related work using machine learning, the accuracy achieved in this chapter could not be compared to any previous work.

The outcome of this work is significant for enabling physicians to make preventative intervention decisions in order to successfully manage the risk of high level HbA1c. The replication of this work using other hospital clinical datasets can ultimately help provide improved healthcare services to patients and reduce the cost and time needed to assess the HbA1c test. This can help identifying patients who are at a high risk of developing T2DM, and has the potential to be used as an early warning indicator for developing serious health complications.

We introduced here the collaborative denoising autoencoders (Col-DAE). The framework uses denoising autoencoders to model separately the high and low level HbA1c data. The latent spaces of both models are then merged and passed to a MLP model for decision making. This framework was utilised for a complex classification challenge (HbA1c level prediction from routinely collected

clinical data) and has shown very promising results. The framework presented here is a general framework and can be generalised for other applications and with multi-class problems.

Epilogue

In this chapter, we have introduced a novel approaches for predicting patients with diabetes levels of Glycated Haemoglobin (HbA1c) using state-of-the-art machine learning methods. The work in this chapter demonstrated that the presence of high HbA1c levels in patients can be reliably predicted from routinely collected clinical data. The following chapter will study predicting Glycated Haemoglobin (HbA1c) elevation levels for patients (with pre-diabetes) who have no history of hyperglycemia.

Chapter 7

Elevated Glycated Haemoglobin Levels Prediction Using Machine Learning for Pre-diabetes Identification

Prologue

Building on the previous work in Chapter 6, we proceed to study current Glycated Haemoglobin (HbA1c) elevation levels for patients with no history of hyperglycemia (to identify patients with pre-diabetes). This chapter is aimed at providing an investigation about using machine learning for the predictability of current Glycated Haemoglobin (HbA1c) elevation levels from the data stored in the electronic health records (EHR).

This chapter sought to address the following research questions (mentioned in the Introduction chapter):

- How can machine learning assist with the early identification of patients with pre-diabetes via the prediction of elevated Glycated Haemoglobin (HbA1c) in patients with no history of hyperglycaemia?
- Can the use of temporal (time-series) data available in EHR help improve the elevated Glycated Haemoglobin (HbA1c) levels prediction using machine learning?
- What is the impact of including temporal behaviour metrics on the importance of variables used to predict elevated Glycated Haemoglobin (HbA1c) levels?

Declaration: Description of these studies as presented in this chapter is largely as published or submitted on the following publications:

Alhassan Z, Budgen D, Alshammari R, Al Moubayed N. *Predicting Current Glycated Haemoglobin Levels in Adults From Electronic Health Records: Validation of Multiple Logistic Regression Algorithm.* JMIR medical informatics. 2020;8(7):e18963 (Alhassan, Budgen, Alshammari and Al Moubayed 2020).

Alhassan Z, Watson M, Budgen D, Alshammari R, Alessa A, Al Moubayed N. *Improving Current Glycated Hemoglobin Prediction in Adults: Use of Machine Learning Algorithms With Electronic Health Records.* JMIR medical informatics.2021;9(5):e25237 (Alhassan, Watson et al. 2021).

The references and notations have been altered, cross-references have been added and some stylistic changes have been made for the consistency throughout this thesis.

7.1 Introduction

The aim of this part of the thesis is to investigate state-of-the-art machine learning approaches for predicting the elevation levels of HbA1c for non-diabetic patients using data that are routinely collected and stored in the EHR systems. As stated in section 2.2 of healthcare context chapter (Chapter 2), patients with an HbA1c level of 5.7% or more are considered to have an elevated HbA1c and those with lower levels than that are considered normal. The overall methodology used to achieve the objectives of this part of the thesis is as follow:

- First, we start with performing a differentiated replication study to validate, evaluate, and identify the strengths and weaknesses of the predictive models to forecast the levels of HbA1c using different (KAIMRC) EHR data.
- Second, we investigate improving the results by employing advanced machine learning approaches, incorporating knowledge extracted from KAIMRC EHR longitudinal data and discuss the relative importance of the predictors. We also investigate the performance of the machine learning models by incorporating the top available predictors extracted from KAIMRC EHR data.

To fulfil our research questions outlined in this thesis, we require a newer and richer dataset (KAIMRC part 2). The utilisation of extra features and larger sample size were important for the robustness and consistency of the objectives to be achieved in this chapter.

7.2 Differentiated Replication Study for Predicting Current HbA1c Elevation Levels in Adults From EHR Data

Many studies have investigated the correlation between HbA1c and clinical variables using statistical and mathematical approaches (McCarter, Hempe and Chalew 2006; Nathan et al.

2008; Kazemi et al. 2014; Rose and Ketchell 2003). However, we are not aware of any that have performed replications of the predictive models on different populations. In this section, we investigate building statistical models that predict the probability of patients having an elevated level of HbA1c. We employ comparative statistical models similar to the models used by Wells, Lenoir et al. (2018) and apply them to a larger electronic health record (EHR) dataset extracted from the KAIMRC EHR systems in Saudi Arabia.

The work by Wells et al., which we refer to in this work as the original study, focused on predicting the level of HbA1c for patients who were not previously diagnosed with diabetes or taking diabetes medications. The data were extracted from the EHR database of Wake Forest Baptist Medical Centre in the United States. The authors applied a multiple logistic regression model to create a mathematical equation for calculating the level of HbA1c (>= 5.7). The predictors used in the equation were chosen from a list of theoretically associated hyperglycemia variables (laboratory measurements, medication categories, diagnosis, vital signs, demographics, family history, and social history variables). After reducing the model's variables using Harrell's model approximation method (Harrell Jr, K. L. Lee and Mark 1996) and removing variables that caused collinearity, the final equation associated eight independent variables with the result of the HbA1c blood test. Restricted cubic splines (RCS) with three knots were used for fitting the continuous predictors into the model (Wells, Lenoir et al. 2018). The calculator achieved an accuracy of 77%.

The independent replication of empirical studies is widely regarded as being an essential underpinning of the scientific paradigm (Walker, James and Brewer 2017). Successful replication of a study by other researchers is considered to be an important step in verifying the original findings and helping to determine how widely they apply (Nosek and Errington 2020; Da Silva et al. 2014).

While the vocabulary associated with replication varies across disciplines (Gómez, Juristo and Vegas 2010), the terms employed by Lindsay and Ehrenberg (1993) appear to be widely used and recognised, so they will be used in this chapter. Lindsay and Ehrenberg categorise replication studies as either (i) close replications or (ii) differentiated replications.

First, a close replication seeks to repeat the original study in a way that keeps all the "known conditions of the study the same or very similar" (Lindsay and Ehrenberg 1993). Hence, such a study employs the same forms of measurement, sampling, and analysis as the original, while also seeking to keep the profile of any set of participants as close to the original as possible. A close replication aims to test the hypothesis that, when a given study is repeated under the same experimental conditions as the original study, it should produce the same (or nearly the same) result.

Second, a differentiated replication introduces known variations into what Lindsay and Ehrenberg term "fairly major aspects of the conditions of the study" (Lindsay and Ehrenberg 1993). Differentiated replications provide a test of how widely the original findings can be generalised, their scope, and the conditions under which they may not hold. For a differentiated replication, therefore, it is expected that some changes in the outcomes are likely to arise, and the question of interest is to what extent and in what form these outcome changes occur.

In an ideal situation, one or more close replications would be used to validate the findings of an original study, followed by a set of differentiated replications used to scope out the extent of their validity by varying different conditions.

For any replication study, it is possible to vary one or more factors from those factors that characterise the way that the study was performed. These may include the team performing the replication, the analysis process, the type of data employed, and the population from which the data were derived. As this study involves analysing data collected from a human population rather than conducting an experiment or trial, we can expect that using a different team to perform a replication should have no effect. Hence, for a close replication it would be appropriate to use the same analysis tool with EHRs of the same form as used in the original study, but pertaining to a different sample of participants drawn from the same general population used in the original study.

For the differentiated replication reported here, we have used the same form of analysis, but have applied this to a set of EHRs that were derived from a different population. The differences

between the forms of the EHRs constituted one difference, but these differences were relatively small. The main difference in the studies arose from the population used. As with the original study, the selection of participants was largely driven by availability. We therefore expected that it was quite possible that there would be some differences in the outcomes, and our main goal was to investigate the extent and form of those differences.

7.2.1 Methodology Employed for the Replication

Conduct of the Replication Study

The employed dataset was extracted from KAIMRC EHR dataset. The dataset records were then labelled according to the ADA guidelines. Patients with an HbA1c level of 5.7% or more are considered to have an elevated HbA1c and those with lower levels than that are considered normal. The predictors that were selected by the authors of the original study for calculating the level of HbA1c, listed in Table 7.1, were employed in this study, except for race and smoking status. Taking into account that most of the data samples in the KAIMRC dataset are from the same race, the race variable can be omitted, as it has zero variance (Austin and Steyerberg 2012). Smoking status information is absent from the KAIMRC dataset. However, in the original model used by Wells et al., this was ranked as having the lowest importance of all the predictors. The BMI and non–high-density lipoprotein measures were also absent. However, both can be calculated by using the formulae presented below[1]:

$$BMI = Weight(kg)/(Height(m))^2 \tag{7.1}$$

$$non\_HDL = CHOL(mmol/L) - HDL(mmol/L) \tag{7.2}$$

[1] Reference: https://www.whittington.nhs.uk/document.ashx?id=10724 and https://www.thecalculatorsite.com/articles/health/bmi-formula-for-bmi-calculations.php respectively.

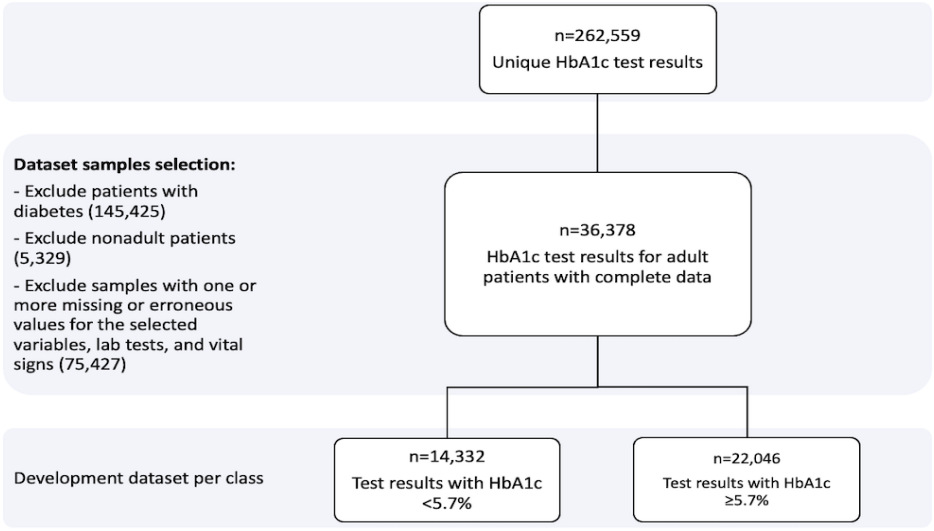
Table 7.1: Predictors available in the original study versus KAIMRC data subset(B).

Predictors	Original study dataset	KAIMRC dataset
Age	✓	✓
Body mass index	✓	✓(calculated)
Estimated glomerular filtration rate	✓	✓
Random blood sugar (glucose) level	✓	✓
Non–high density lipoprotein	✓	✓(calculated)
Total cholesterol	✓	✓
Race	✓	✗
Smoking status	✓	✗

In this study we followed the same sampling approach used in original study. For inpatient visits, only the first day's data were considered, and in cases of missing values, the first available values for the visit were used. Samples for patients with values of <1% for HbA1c were simply considered to be erroneous readings and were excluded. Similar to the original study, patients diagnosed with diabetes were eliminated from the development dataset (refer to Appendix C for diabetes diagnostic codes). We avoided intensive interpretation for handling the missing values. Samples with one or more completely missing values were also excluded. This resulted in decreasing the dataset, KAIMRC data subset(B), size from the 262,559 samples originally collected to 36,378 samples. Figure 7.1 shows the detailed preprocessing tasks performed prior to building the statistical models. Refer to Appendix B for more details (ICD10 code and description) about the variables employed in this data subset.

The descriptive statistics for the KAIMRC experimental dataset and the dataset used by Wells et al. are shown in Table 7.2. The units used for recording lab tests can differ according to the laboratory guidelines followed by each country. The KAIMRC dataset uses different units than the ones used in the original study for some variables. For instance, the total cholesterol level is measured in milligrams per deciliter (mg/dL) in the original study's dataset, and in millimoles per liter (mmol/L) in the dataset from the KAIMRC labs. Therefore, the descriptive statistics contain the values using both units. When developing the predictive models, we converted the units using the appropriate formulae, 7.1 and 7.2 respectively (previously provided in subsection 7.2.1 of this chapter). However, the conversion task can be avoided to reduce data preprocessing complexity, as it should not affect the prediction performance for the logistic

Figure 7.1: Dataset preprocessing details used for the replication.



regression models (this was experimentally checked by redoing the experiments with and without the unit conversion of the values for the variables used).

Table 7.2: Descriptive statistics of the selected features in the KAIMRC data subset(B).

Feature	**Unit**	**KAIMRC dataset**			**Original study dataset**		
		HbA1c <5.7%	**HbA1c >=5.7%**	*P* **value**	**HbA1c <5.7%**	**HbA1c >=5.7%**	*P* **Value**
Age, Mean (SD)	Years	45.5 (17.01)	60.5 (14.13)	<0.001	48.1 (15.4)	54.8 (14.0)	<0.001
BMI, Mean (SD)	kg/m^2	29.26 (6.77)	30.94 (6.31)	<0.001	30.1 (7.44)	33.0 (8.41)	<0.001
eGFR: Estimated Glomerular Filtration Rate, Mean (SD)	mL/min/1.73 m2	93.40 (35.19)	82.02 (28.86)	<0.001	92.0 (33.0)	87.9 (30.8)	<0.001
RBS: Random Blood Sugar (glucose), Mean (SD)	mmol/L	5.47 (1.28)	8.30 (4.30)	<0.001	4.9 (0.7)	5.3 (0.9)	<0.001
	mg/dL	98.5 (23.00)	149.4 (77.47)		88.4 (12.7)	96.1 (16.0)	
CHOL: Total cholesterol, Mean (SD)	mmol/L	4.59 (1.19)	4.17 (1.16)	<0.001	4.80 (1.01)	4.96 (1.11)	<0.001
	mg/dL	177.49 (46.01)	161.25 (44.85)		186 (39.4)	192 (43.1)	
non-HDL: non-High Density Lipoprotein, Mean (SD)	mmol/L	2.85 (1.06)	2.49 (0.99)	<0.001	3.49 (0.96)	3.72 (1.07)	<0.001
	mg/dL	110.2 (40.99)	96.28 (38.28)		135 (37.4)	144 (41.7)	

Study Design

A complete validation of Wells et al.'s calculator using our dataset was not possible due to the absence of the smoking status variable. To validate the approach used in the original study, 3 predictive models (PMs) were built, trained, and tested using the KAIMRC data subset(B). All models employ multiple logistic regression to create the calculator by associating the chosen and available predictors. After discussion with the authors of the original study, we structured the models as PM1, PM2, and PM3.

- PM1 was designed to be as close as possible to the original study's model. It uses the predictors chosen in the original study: age, BMI, Random Blood Sugar (RBS), non–high-density lipoprotein (non-HDL), cholesterol, and estimated Glomerular Filtration Rate (eGFR). The continuous predictors are fitted to the model using RCS with 3 knots.

- PM2 was designed using the same predictors used in PM1 but without RCS fitting.

- PM3 was designed after excluding the predictors with the least importance in PM1 and PM2, using a reduced number of predictors and fitted using RCS with 5 knots. The choice of the number of knots for this model was determined by using Stone's recommendation (Austin and Steyerberg 2012).

The 3 models were validated using the 10-fold cross-validation approach. The measure used to evaluate and compare the results with the original study was the concordance statistic, which is equal to area under the receiver operating characteristic (AUR-ROC) curve (Austin and Steyerberg 2012). To assist with future comparisons, we report measures commonly used for medical research, such as precision, recall, and F1, in the model evaluation. The data preparations are undertaken using Python (version 3.7; Python Software Foundation). The model building and the analysis are carried out in R (version 3.6.0; The R Foundation) using the regression modelling strategies package.

7.2.2 Results

The development data subset size used for training, testing, and validating the models after data preprocessing was 36,378 samples. Most medical datasets are imbalanced with a majority normal population (Saito and Rehmsmeier 2015), but 60.60% (22,046/36,378) of KAIMRC dataset patients were found to have elevated levels of HbA1c (>=5.7%), and 39.40% (14,332/36,378) of patients had a normal HbA1c level (<5.7%).

Details of the 3 models (PM1, PM2, and PM3) used for the purpose of validating and evaluating the original study are shown in Table 7.3. This study explores multiple logistic regression models using different numbers of variables, with and without RCS, and with different numbers of knots. PM1 (using a complete set of variables fitted using RCS) achieves an average accuracy of 73.67% and 95% CI of 74% to 77% with a well-calibrated curve. A similar model (PM2), but not fitted using RCS, shows improved accuracy, with an average accuracy of 74.04% and the same 95% CI of 74% to 77%. However, the calibration curve shows better calibration when applying RCS into the models, as shown in Figures 7.2 and 7.3.

Table 7.3: Performance of models for HbA1c elevation prediction.

Model	Variables used	Number of RCS knots	AUR ROC	95% CI	Recall	Precision	F1
PM1	Complete	3	73.67	74.71-77.51	85.24	77.58	81.23
PM2	Complete	N/A	74.04	74.35-77.16	82.18	78.76	80.43
PM3	Reduced	5	74.73	75.38-78.15	84.40	78.80	81.50

Figure 7.4 shows the ranking of importance for the variables used in the PM1 model. PM1 shows a different order of importance for the predictors than the order obtained from the original study. Age and RBS are of great importance in both studies. However, BMI is of the lowest importance when using the KAIMRC population, whereas in the original study it was ranked second.

The PM3 model excludes the variable that showed the lowest importance, BMI. This model, when fitted using RCS with 5 knots, shows better performance using only the 5 predictors (age, RBS, cholesterol, eGFR, and non-HDL). The eGFR shows greater importance when fitted using RCS with 5 knots (>0.05) than when fitted with 3 knots (<0.05). The predictors' importance

Figure 7.2: The calibration curve for PM1.

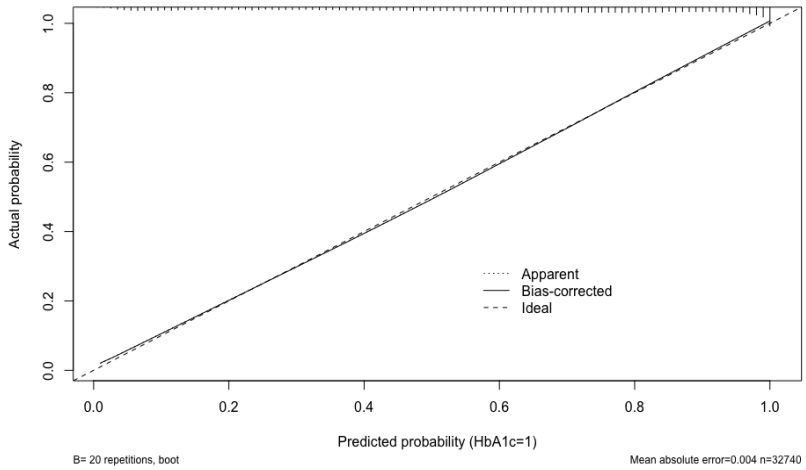


Figure 7.3: The calibration curve for PM2.

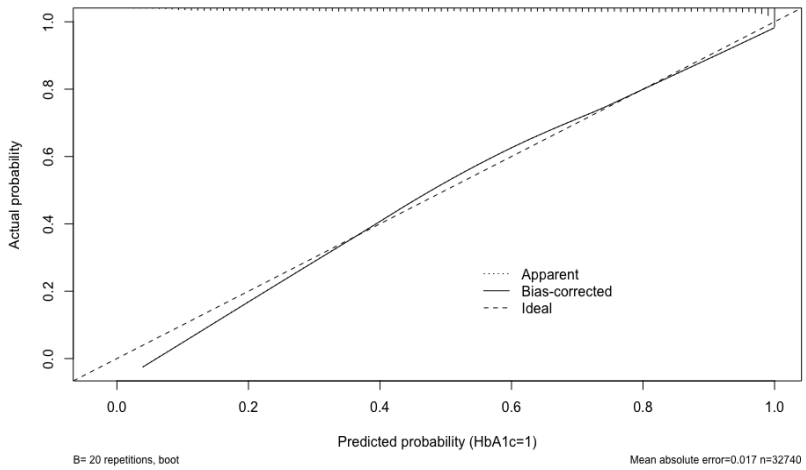
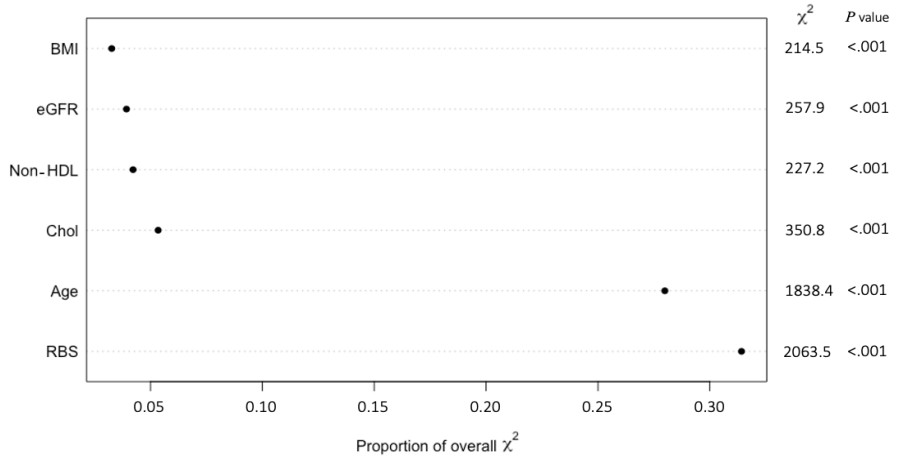


Figure 7.4: Order of importance of predictors for PM1.



order for PM3 is shown in Figure 7.5. PM3 achieves an average accuracy of 74.73%, with a better confidence interval (95% CI 75%-78%). The calibration curve for PM3 is identical to that of PM1.

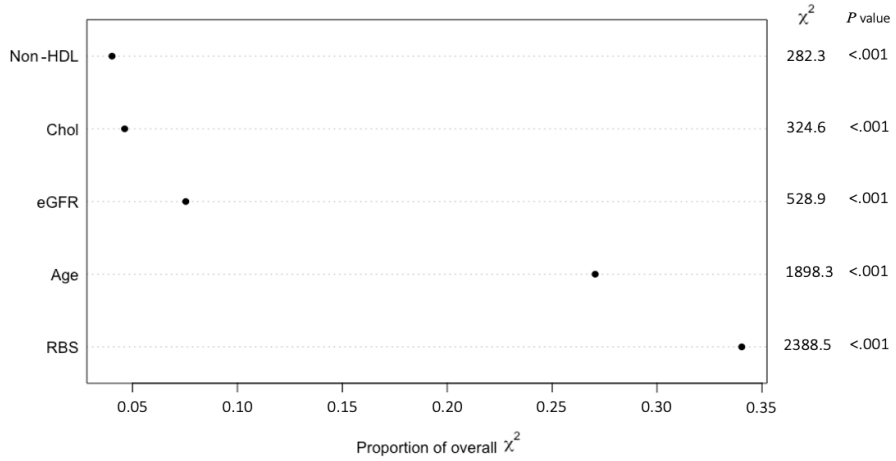
When using the PM2 model, the results show agreement with the results from PM1 for 93.27% (33,929/36,378) of predictions. The PM3 model with fewer predictors achieves a better performance and a similar percentage of predictions that are in agreement with the output from PM1 (33,937/36,378, 93.29%). Furthermore, the results show a strong degree of correlation among the probability outputs produced by the 3 models ($r = 0.97$).

7.2.3 Discussion and Conclusion

Principal Results

Applying the method employed in the original study achieved an accuracy of 73% to 74% using a dataset collected from the Middle East, compared with 77% obtained from using a population from the United States in the original study. The findings from this replication study

Figure 7.5: Order of importance of predictors for PM3.



therefore confirm the conclusion from the original study that this form of modelling can help with predicting the levels of HbA1c in a blood test for non-diabetic patients using predictors extracted from EHR systems.

The order of importance obtained for the predictors used by the multiple logistic regression on our dataset is different from the order of importance produced in the original study. The order for the predictors using the KAIMRC dataset, from the most to the least importance, is RBS, age, eGFR, cholesterol, non-HDL, and BMI. Table 7.4 shows the importance rankings for the predictors obtained from the original study, as well as the rankings obtained from the 3 models used in this study.

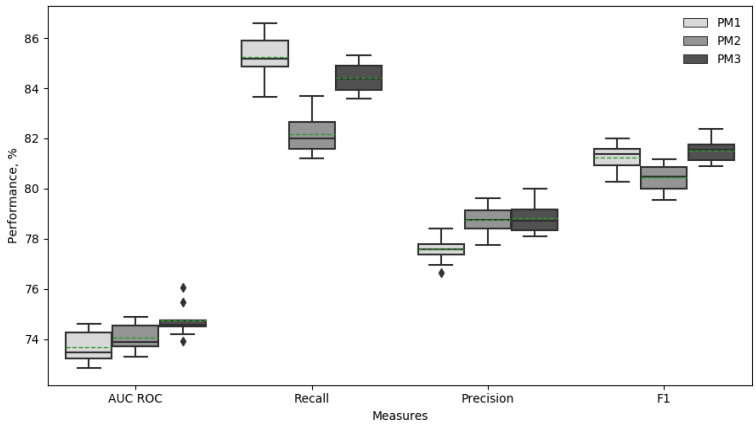
BMI was one of the most important predictors in the population from the United States and demonstrated higher impact than the RBS and eGFR. However, it shows little importance for predicting the elevation level of HbA1c in the KAIMRC population. Indeed, the simpler calculator with a reduced number of variables (after excluding BMI) is able to achieve better prediction abilities (refer to Appendix E for details of the calculator). Figure 7.6 summarises

Table 7.4: Predictors importance rankings obtained for the PMs used.

Study		1st	2nd	3rd	4th	5th	6th	7th	8th
Original study		Age	BMI	RBS	Race	Non-HDL	Cholesterol	eGFR	Smoking status
Replication study	PM1	RBS	Age	Cholesterol	Non-HDL	eGFR	BMI	N/A	N/A
	PM2	Age	RBS	Cholesterol	Non-HDL	BMI	eGFR	N/A	N/A
	PM3	RBS	Age	Cholesterol	Non-HDL	eGFR	BMI (excluded)	N/A	N/A

the 10-folds performance achieved using the reported measures for all models, and reveals that there is a consistent prediction trend for PM3, especially in the AUR ROC, which shows little variation between the folds.

Figure 7.6: Box plots of the AUC-ROC, recall, precision and F1 measures performance for the PMs used.



This replication study shows that the ranking of the variables is largely based on the dataset and the model used for prediction. Variables with low importance in the prediction of HbA1c in one population may show greater or lesser importance when the model is applied on populations from different regions of the world. Interestingly, this can also happen when employing different predictive models and with different hyperparameters using the same population (for instance, eGFR shows higher importance when fitted to the model using RCS with 5 knots in PM3 than

with 3 knots in PM1 and without RCS in PM2, as interpreted in Table 7.4.

Limitations

We performed a differentiated replication using a population from a different region that was available to us. The two datasets have similar means and standard deviations for most of the variables, such as age, cholesterol, and non-HDL, as described in Table 7.2. However, there is a difference in the body mass index and random blood sugar variables, and the dispersion is large for both variables.

The sample size and class balance affect the learning behaviour of the models (Batista, Prati and Monard 2004). The KAIMRC dataset is larger than the one used in the original study by 38%. The class balance is also different, with 26% of patients having elevated HbA1c (>=5.7%) and 74% with normal HbA1c (<5.7%) in the original study compared with 60.60% (22,046/36,378) with elevated HbA1c (>=5.7%) and 39.40% (14,332/36,378) with normal HbA1c (<5.7) in KAIMRC dataset. Also, the sampling approach used by Wells et al. is a visit based sampling. This means that the data used for training and testing the models can contain multiple samples (visits) made by same patient.

Although the population represented in this study is less heterogeneous with regard to ethic groups, the size of the KAIMRC dataset is larger than the one used in the original study. The prevalence of diabetes is also larger, being a sample from the population of Saudi Arabia. In terms of prevalence of diabetes, Saudi Arabia was ranked by the World Health Organisation as being the second highest in the Middle East and seventh highest in the world (Abdulaziz Al Dawish et al. 2016), with an 18.3% diabetes prevalence rate, according to the IDF, compared with 10.5% in the United States (Centers for Disease Control and Prevention 2020).

In the original study, the model performance was compared with the models developed by Baan et al. (1999) and Griffin et al. (2000), which used different datasets (D. Williams et al. 1995; Kinmonth, Spiegal and Woodcock 1996). The main limitation in the comparison between the original study and the studies by Baan et al. and Griffin et al. is the absence of some variables

that were used to create the calculators (refer to Table 7.5 for details about the variables used in the corresponding studies and and this study). The same situation applies to this study, as the smoking status variable is missing in the KAIMRC dataset. The smoking prevalence in Saudi Arabia is between 2.4% to 52.3% among different age groups (Bassiony et al. 2009). However, other missing predictors, such as genetic or lifestyle characteristics (Elhadd, Al-Amoudi and Alzahrani 2007), which are difficult to collect and incorporate into the EHR systems, may help to explain the high rate of elevated levels of HbA1c in the KAIMRC population.

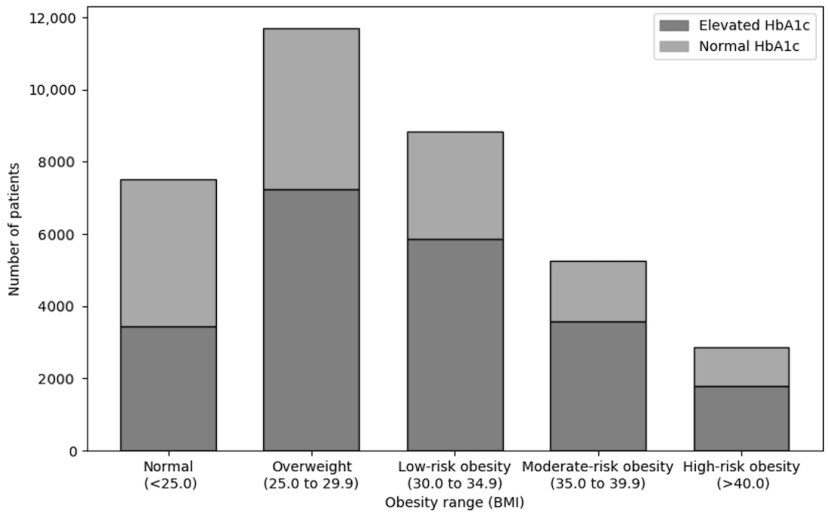
Table 7.5: The variables used in the comparative studies investigated in the replication study.

Predictors	Wells, Lenoir et al. (2018)	Baan et al. (1999)	Griffin et al. (2000)	Our study
Age	✓	✓	✓	✓
Gender	✗	✗	✓	✗
Height	✗	✗	✗	✗
Weight	✗	✗	✗	✗
Body Mass Index (BMI)	✓	✓	✓	✓
Gestational diabetes	✗	✓	✗	✗
Prevalence of cardiovascular disease	✗	✓	✗	✗
History of diabetes	✗	✗	✓	✗
Prescribed antihypertensive medication	✗	✓	✓	✗
Prescribed lipid-lowering medication	✗	✓	✓	✗
Prescribed steroids medication	✗	✗	✓	✗
estimated Glomerular Filtration Rate (eGFR)	✓	✗	✗	✓
Random Blood Sugar (RBS)	✓	✗	✗	✓
Non-High Density Lipoprotien (non-HDL)	✓	✗	✗	✓
Total Cholesterol (CHOL)	✓	✗	✗	✓
Race	✓	✗	✗	✗
Smoking status	✓	✗	✓	✗
Physical Activities	✗	✗	✓	✗

After eliminating the variables that do not show significant impact on the prediction of HbA1c in the KAIMRC population, the results indicate that different regions in the world can have different weightings of predictors for HbA1c when using the approach of Wells et al. Although there are many studies that have demonstrated the relationship between diabetes prevalence and BMI (Boffetta et al. 2011), some studies have shown that the obesity prevalence in Asian countries does not relate to the diabetes prevalence. The risk of diabetes occurs in patients with a lower BMI in Asian countries compared with patients from European countries (Yoon et al. 2006). The prevalence of obesity in Asian countries is substantially less than in the United

States, but Asian countries have a similar or higher prevalence of diabetes (Hu 2011). However, neither Yoon et al. (2006) nor Hu (2011) identifies a relationship between non-diabetic patients with elevated levels of HbA1c and obesity. Figure 7.7 visualises the class distribution for the BMI variable for the KAIMRC dataset. The figure shows that elevation of HbA1c exists with similar rates between low and high obesity ranges.

Figure 7.7: HbA1c elevations for BMI ranges of King Abdullah International Medical Research Center patients using data subset(B).



Conclusions

Replication studies provide an invaluable contribution to the validation, generalisation, and continuation of scientific research. The differentiated replication presented in this study is aimed at validating the calculator used for predicting HbA1c and evaluating the method used to create the mathematical equation by training the multiple logistic regression algorithm using EHR datasets. The evaluation was performed using a dataset collected from a different population. The original and replicated calculators employ associated predictors that are routinely collected and stored in hospital systems.

This differentiated replication study used the same method to analyse a different population sample, with some differences in the form of the EHRs. As a replication, it was intended to investigate what changed and did not change in the outcomes.

What did not change appreciably was the accuracy of the results produced using this method, with an accuracy range of 73.6% to 74.7% in our study compared with 77% in the original study. The set of predictors (when these could be compared) also did not change. Thus, given that a close replication of the original study is unavailable, the differentiated replication does confirm that, despite the notable differences between the two datasets, the use of multiple logistic regression is able to provide good predictions of HbA1c elevation levels.

What did change was the order of importance for the set of predictors used in the calculator. Thus, we can conclude that the use of multiple logistic regression for prediction does need to be tuned to the characteristics of the population being assessed. While we cannot wholly rule out the cause of this difference in importance being due to differences in the form of the EHRs, it seems more likely that the characteristics of the population were an important factor.

In terms of the role of replication itself, we would argue that this study demonstrates that while there is little difference in prediction accuracy when using multiple logistic regression with different populations (as might be expected), the influence of the different elements in the set of predictors is different. Due to that, we would argue that the generalisation of simple statistical predictive models (calculators) is inappropriate. We suggest that creating advanced predictive models that can learn complex relationships using large multidimensional datasets may be a better way to exploit the increasing volumes of EHR data becoming available. Building on this work, next section will investigate applying advanced machine learning techniques and the use of longitudinal data to predict the elevation of HbA1c.

7.3 Current Glycated Haemoglobin Prediction in Adults: Use of Explainable Machine Learning Algorithms with Longitudinal Data from Electronic Health Records

In this section of this chapter, we investigate the performance of predictive models to forecast HbA1c elevation levels (for patients with no history of hyperglycemia) by employing machine learning. We also aim to investigate utilising the patient's EHR longitudinal data in the performance of the predictive models.

First, this section reports an investigation into the performance of machine learning models to predict current HbA1c levels using the features used in the replication study (section 7.2). Second, in addition to the features used in the replication study, we also investigate the performance of the models using those features along with the top available features in KAIMRC dataset.

7.3.1 Method

To study the impact of employing advance machine learning predictive models on the predictability of patients' current HbA1c levels from EHR data, we first employed the Multiple Logistic Regression (MLR), used in the replication study (for comparison). We also employed commonly used (based on the literature review presented in chapter 4) machine learning models; Random Forest (RF), Support Vector Machine (SVM) and Logistic Regression (LR) models; as well as a deep learning model, Multi-layer Perceptron (MLP) (LeCun, Bengio and Hinton 2015)).

The problem was formulated into binary classification problem whereby the target variable, HbA1c level, was encoded with one when the level of HbA1c is 5.7% or more and with zero otherwise. The results obtained from using these models were compared to those obtained from employing the model used by Wells, Lenoir et al. (2018) with the KAIMRC dataset (detailed in the Dataset subsection). We also investigated the impact of including longitudinal patient data upon their performance. The performance of each model was evaluated using measures commonly

employed in clinical applications. For the SVM and MLP models, the relative importance of the features was also calculated using explainable machine learning techniques.

Using black box machine learning models in healthcare can have adverse effects on the trust and confidence in the outcomes (Ahmad, Eckert and Teredesai 2018). Explainable methods for machine learning models allow interpretable outcomes that can expose the reasons behind the decision made by the model (Lipton 2018). This transparency provides both health professionals and patients with the confidence and trust in the outcome of the models. The SHAP (SHapley Additive exPlanations) values (Lundberg and S.-I. Lee 2017) and LIME scores (Ribeiro, Singh and Guestrin 2016) techniques have therefore been employed to provide a degree of transparency to our deep learning models.

SHAP values are derived from Shapley values used in game theory, and provide a method of calculating the contribution of each feature to the final prediction via the GradientSHAP approximation. This is achieved for each feature by comparing the prediction the model makes when the feature is present with the prediction obtained when the feature takes some baseline value (Lundberg and S.-I. Lee 2017). Consequently, the SHAP values for a given input 'explain' how each feature affects the output of the model when compared to the baseline (or 'default') output of the model. We use SHAP values to interpret our black box models as they can be efficiently calculated, and allow a global view of the model to be constructed through the computation of SHAP values from across the whole dataset.

SHAP values are computed using the feature's mean marginal contribution across different coalitions of all features. Shapley values themselves are computationally intensive to compute, and so approximation methods are commonly used when calculating the values.

To ensure that the SHAP values we calculate are not too greatly affected by the approximation method used, we also compute the LIME (Ribeiro, Singh and Guestrin 2016) scores for the models, across the entire dataset. LIME tries to estimate locally faithful linear explanations (i.e. explanations that correspond to how the model behaves around the instance being explained)

for any classifier. LIME achieves this by creating local linear classifiers that approximate the behaviour of the original model in the vicinity of the data being explained. As linear models are inherently interpretable through their parameters, they can be used to generate explanations of the original model. Both SHAP and LIME have the advantage that they are model-agnostic techniques, and so we are able to apply both methods to all of our black box classification models (Lundberg and S.-I. Lee 2017; Ribeiro, Singh and Guestrin 2016).

Dataset Preparation

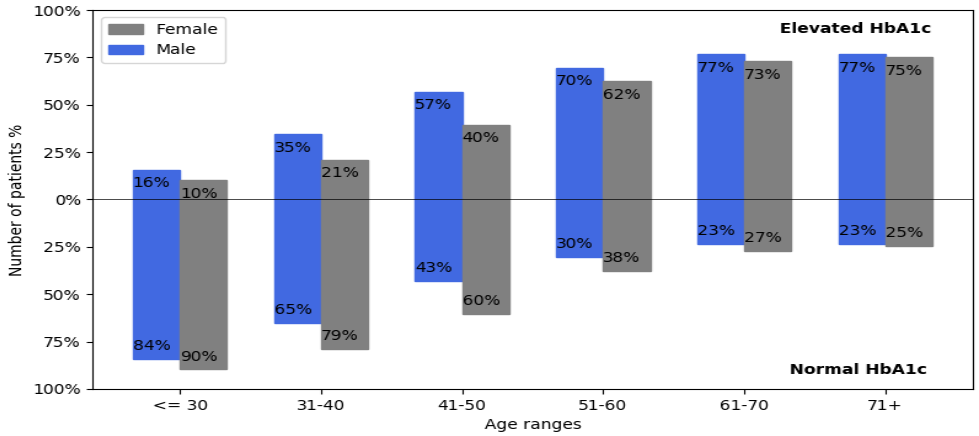
The dataset contains a full history of patient details, vital signs, and lab test readings for each patient visit for the period from 2016 to the end of 2018. As the aim of this study is to identify non-diabetic patients that are at a high risk of HbA1c elevation, all patients previously diagnosed with hyperglycemia were eliminated from the experimental dataset. The remaining cohort formed our experimental dataset.

Most medical datasets are imbalanced (Batista, Prati and Monard 2004; L. Zhang, H. Yang and Jiang 2018; Rahman and Davis 2013). Such imbalances occur when the proportion of one class of patients in the dataset is greater than its counterpart class (Longadge and Dongre 2013). However, unusually, our experimental dataset is not imbalanced. Slightly over half of the patients in our experimental dataset (52.1%) were found to have elevated levels of HbA1c (>=5.7%) while 47.9% of patients had normal HbA1c levels (<5.7%). This can be ascribed to the high incidence of diabetes in the region from which the dataset was collected (Alqurashi, Aljabri and Bokhari 2011).

A detailed illustration of the patients' class distribution (HbA1c levels) by age groups and gender is shown in Figure 7.8. Figure 7.8 shows that as the age of patients increases, the proportion of patients who have elevated HbA1c levels is steadily increasing.

The dataset exhibits a balanced gender distribution with 49.4% of the patients were male and 50.6% female. However, the proportion of male patients with elevated levels of HbA1c (>=5.7%) is more than for the female patients. Also, female patients with normal levels of HbA1c (<5.7%)

Figure 7.8: HbA1c Elevation levels distributed over age range and gender in the KAIMRC dataset (before sampling).



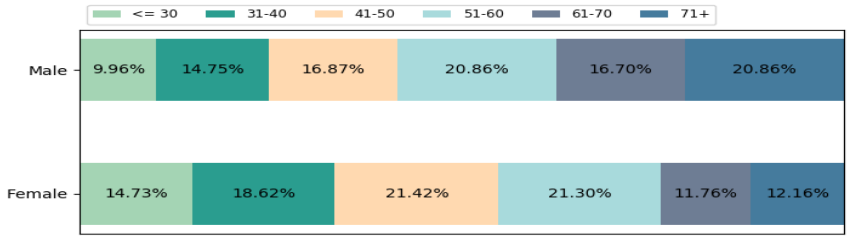
made more visits than males. Table 7.6 shows the profile for the distribution of HbA1c elevation levels organised by gender.

Table 7.6: Profile for the class distribution over gender.

Characteristics		**HbA1c <5.7%**	**HbA1c >=5.7%**
Number of patients (Total:18,844)	Total (%)	9,018 (47.9%)	9,826 (52.1%)
	Male%	41 .73%	56.42%
	Female%	58.27%	43.58%
Number of visits (Total: 157,600)	Total (%)	79,607	77,993
	Male%	39.72%	53.32%
	Female%	60.28%	46.68%

The dataset also exhibits a balanced gender distribution with 49.4% of the patients were male and 50.6% female. The number of patients by classified age groups for each gender is shown in Figure 7.9.

Figure 7.9: Number of patients by age groups and gender. Total number of patients is: 18,844 (9,308 male and 9,536 female).



Feature Selection and Data Sampling

First, six main variables (those used in the replication study) were originally extracted from the KAIMRC EHR dataset to be used in this study. The data subset is indexed as data subset(C). The features were selected firstly for their theoretical association with hyperglycemia based on the study by Wells, Lenoir et al. (2018) and secondly for their availability in the KAIMRC dataset. The selected features are: Age, Body Mass Index (BMI), Estimated Glomerular Filtration Rate (eGFR), Random Blood Sugar (RBS), Total cholesterol (CHOL) and non-high density lipoprotein (non-HDL). Refer to Appendix B for more details (ICD10 code and description) about the variables employed in data subset(C). The descriptive statistics (using the data for the current visit only for unique patients), units, and P values for the selected features are presented in Table 7.7.

Table 7.7: Descriptive statistics of the selected features from the KAIMRC data subset(C).

Characteristics	**Unit**	**HbA1c <5.7%**	**HbA1c >=5.7%**	***P* Value**
Age mean (SD)	Years	43.94 (16.38)	58.92 (15.12)	<0.05
BMI[a] mean (SD)	Kg/m^2	29.11 (6.75)	30.90 (6.55)	<0.05
eGFR[b] mean (SD)	$mL/min/1.73m^2$	100.03 (29.22)	85.81 (28.239)	<0.05
RBS[c] mean (SD)	$mmol/L$	5.45 (1.26)	7.88 (4.19)	<0.05
Cholesterol mean (SD)	$mmol/L$	4.65 (1.07)	4.42 (1.20)	<0.05
non-HDL[d] mean (SD)	$mmol/L$	3.45 (1.01)	3.37 (1.11)	<0.05

Second, we also investigate HbA1c elevation levels prediction with the inclusion of variables that have shown clinical correlations with hyperglycemia such as Fasting Blood Sugar (FBS) (Naqvi et al. 2017) and those that are commonly available in the KAIMRC EHR dataset such as Serum

Creatinine level (Crea) and Haemoglobin (Hgb). The investigation will also include studying the impact of integrating the longitudinal behaviour for those variables.

The variables (lab tests) that are commonly ordered by clinicians along with the laboratory tests originally selected in data subsets(B) and (C), Age, BMI, eGFR, RBS, Chol and Non-HDL, are those that have fewer missing values. For instance, the data subset (C) shows that the Creatinine Level (Crea) blood test, was also ordered for all patients along with those variables that were used in subsets(B) and (C). Figure H.1 in Appendix H visualises the number of patients with missing values for the available variables in KAIMRC dataset. The data subset used for this study, indexed as data subset(D). The variables are FBS in addition to the top 20 available variables

To minimise the effect of the sampling approach used on the population size and for fairer comparison with results using six features (and also the results obtained in section 7.2), the missing values in the included variables are replaced with zero instead of removing the complete record. The descriptive statistics (using the data for the current visit only for unique patients), units, and P values for the added features are presented in Table 7.8.

The table shows that the P values for Mean Cell Volume (MPV), and Carbon Dioxide Level (CO2) variables are greater than (0.05). This suggests that there is no statistically significant difference between the classes of HbA1c in those variables.

For the purpose of the studies in this section we aim at predicting the HbA1c value for current patient visits only. Unlike the sampling approach used by Wells et al., which was based on independent hospital visits for patients (including for the same patients), the sampling approach used in this study includes independent patients, to ensure only unseen patients data are used for testing the models.

Since we aim at identifying patients with elevated levels of HbA1c from non-diabetic population, patients previously diagnosed with diabetes were excluded. We also excluded non-adult patients and those with erroneous or missing values. Figure 7.10 shows the details of the tasks performed

Table 7.8: Descriptive statistics of the added variables in the KAIMRC data subset(D).

Variables	**Unit**	**HbA1c <5.7%**	**HbA1c >=5.7%**	***P* Value**
Fasting Blood Sugar (FBS) mean (SD)	$mmol/L$	4.39 (2.05)	5.87 (4.03)	<0.05
Triglyceride Level (TRIG) mean (SD)	$mmol/L$	1.14 (0.7)	1.42 (0.92)	<0.05
High-Density Lipoprotein (HDL) mean (SD)	$mmol/L$	1.19 (0.32)	1.04 (0.30)	<0.05
Chloride Level, Serum (Cl Level) mean (SD)	mEq/L	105.28 (3.6)	104.29 (3.97)	<0.05
Anion Agap (AGAP) mean (SD)	mEq/L	13.09 (3.30)	13.95 (3.36)	<0.05
Urea Nitrogen, Blood (BUN) mean (SD)	$mmol/L$	5.07 (3.92)	6.40 (4.69)	<0.05
Mean Corpuscular Volume (MCV) mean (SD)	femtoliters fL	87.93 (8.61)	87.46 (9.12)	<0.05
Red Blood Cells Width (RDW) mean (SD)	%	13.75 (1.90)	14.10 (1.94)	<0.05
White Blood Cell Count (WBC) mean (SD)	$x10^9/L$	7.09 (2.85)	7.86 (4.00)	<0.05
Mean Cell Volume (MPV) mean (SD)	femtoliters fL	8.42 (1.30)	8.44 (1.32)	0.06
Red Blood Cell (RBC) mean (SD)	$x10^{12}/L$	4.70 (0.69)	4.75 (0.74)	<0.05
Haematocrit (Hct) mean (SD)	$mmol/L$	0.41 (0.06)	0.41 (0.06)	<0.05
Haemoglobin (Hgb) mean (SD)	g/L	133.40 (21.02)	133.13 (21.83)	0.29
Platelets (PLT) mean (SD)	$x10^9/l$	271.95 (84.88)	271.91 (92.37)	<0.05
Mean Cell Haemoglobin (MCH) mean (SD)	pg	28.31 (3.09)	27.98 (3.24)	<0.05
Mean Cell Haemoglobin Concentration (MCHC) mean (SD)	g/l	320.73 (22.70)	318.43 (23.77)	<0.05
Total Bilirubin (T Bili) mean (SD)	$mmol/L$	11.41 (15.67)	10.97 (13.62)	<0.05
Sodium Level, Serum (Sodi) mean (SD)	mEq/L	137.85 (3.73)	137.44 (3.61)	<0.05
Potassium Level, Serum (K Level) mean (SD)	$mmol/L$	4.27 (0.43)	4.37 (0.47)	<0.05
Creatinine Level, Serum (Crea) mean (SD)	$mmol/L$	94.54 (139.90)	96.12 (96.06)	<0.05
Carbon Dioxide Level (CO2) mean (SD)	$mmol/L$	22.95 (3.12)	22.94 (3.40)	0.33

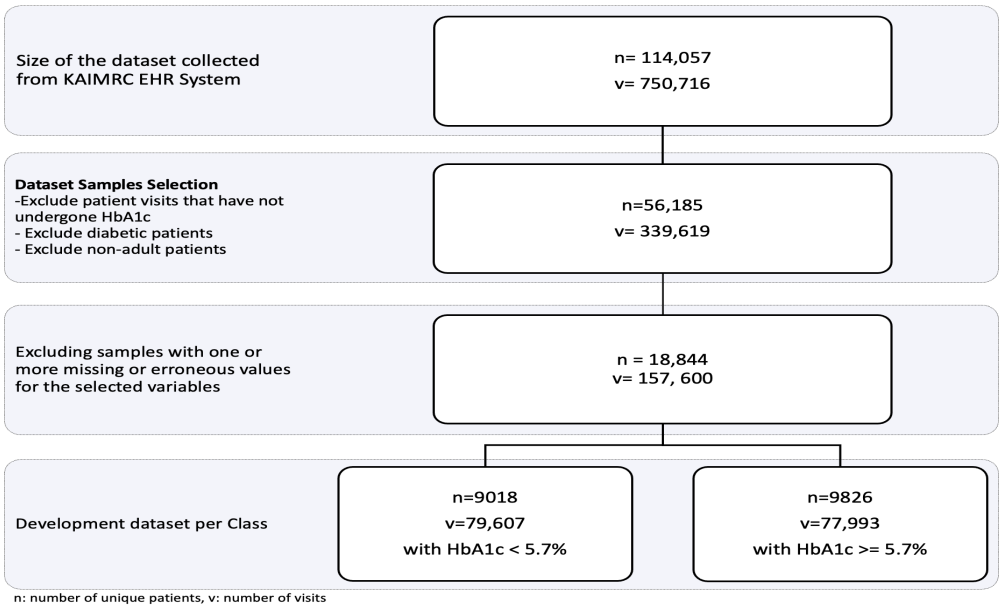
to refine the sample selection. This resulted in a reduction in the size of the experimental dataset from 114,057 patients with 750,709 visits to 18,844 unique patients with 157,600 visits.

Input Preparation for the Models

The input structure for the deep machine learning model was organised as a matrix, based on current and previous time-stamped patient visits. It contained the current visit data concatenated with approximated values for the selected features from all previous visits, which we refer to as the "Approximated Time Series Data".

Each patient visit is described by the selected features, represented as $x_1, x_2, ..., x_n$ Those features are formed as episodes based on the time-stamped values available in each visit (v_i).

Figure 7.10: Details of the sampling approach performed on the KAIMRC dataset.



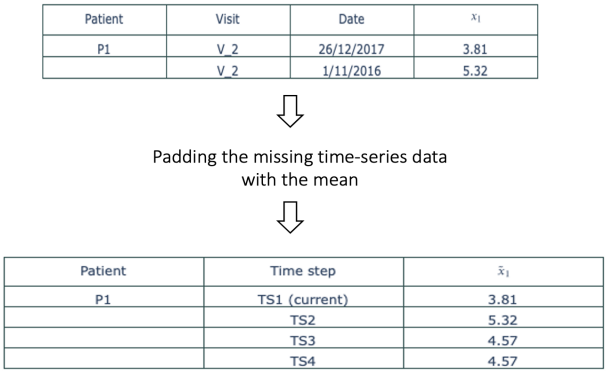
$$Input = \begin{Bmatrix} v_1 : x_{11} & x_{12} & ... & x_{1n} \\ v_2 : x_{21} & x_{22} & ... & x_{2n} \\ ... & ... & ... & ... \\ ... & ... & ... & ... \\ v_s : x_{s1} & x_{s2} & ... & x_{sn} \end{Bmatrix}$$

Here x_{ij} is the feature value at a patient visit v_i ($0 < i >= s$, $0 < j >= n$). s is the number of time series steps (the length of the input sequence) and n is the number of features for each time step, which is set to 6 (or 27 when using the top available features) as explained earlier.

If the number of visits (longitudinal time-series visits) for a patient is less than s, the input for this patient is padded out with the mean value of the available visits to compensate for the missing time-series data. Figure 7.11 shows an example of the padding approach used (in the

figure, we used cholesterol (x_1) as an example feature to demonstrate the padding approach). Where the number of longitudinal visits for a patient is more than s, Piece-wise Aggregation Approximation (PAA) technique (Keogh et al. 2001) is applied to the data for these visits to take account of all data from patient visits.

Figure 7.11: Example of input padding when number of patient longitudinal visits is fewer than s.



PAA transforms the longitudinal time-series data using s as a number of sliding windows (or segments), into a reduced number of time steps data (approximated) employing the mean value of the series falling within that window (segment) (Zhao et al. 2017). We tested the models with several values for the sliding windows (s), 3 was shown to be the optimal value. The formula used to calculate the approximated time-series data is:

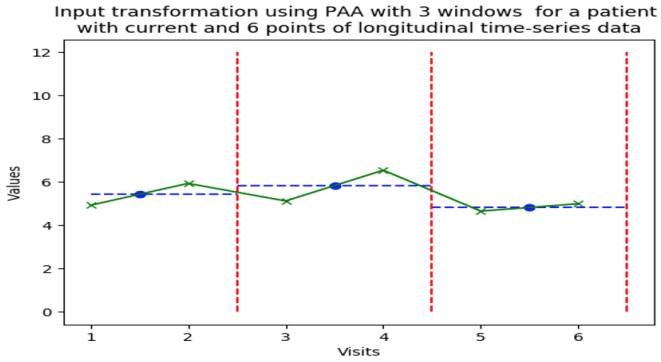
$$\tilde{x} = \frac{s}{r} \sum_{j=(r-\frac{1}{s})(i-1)+1}^{(r-\frac{1}{s})i} x_j, s < r-1 \tag{7.3}$$

where $\tilde{x}_i$ represents the approximated value for x and r is the total number of visits for a patient. s is the reduced number of time-series steps. Figure 7.12 shows an example of the PAA technique using a sliding window of $s = 3$ for Cholesterol (x_1) feature for a patient with current visit and

six time-stamped visits that were available in the EHR longitudinal data for the given patient ($r = 7$).

Figure 7.12: Example of time series steps transformation using PAA for the cholesterol feature when the number of patient visits is more than $s = 3$.

Patient	Visit	Date	x_1
P2	V_1	23/12/2018	4.72
	V_2	5/7/2018	4.93
	V_3	31/1/2018	5.94
	V_4	20/9/2017	5.13
	V_5	23/4/2017	6.55
	V_6	2/10/2016	4.66
	V_7	9/5/2016	5



Patient	Time step	$\tilde{x}_1$
P2	TS1 (current)	4.72
	TS2	5.44
	TS3	5.84
	TS4	4.83

In this example, the number of visits is $r = 7$ (current visit and 6 longitudinal time series visits as shown in top table in the figure). The cholesterol value in the current visit is unchanged. The cholesterol values in the longitudinal time series data (6 visits) are transformed into 3 values using the PAA as shown in the bottom table.

The approximated time-series data forming the output of the PAA is then concatenated with the current visit data to form the final input for the deep learning model. Since the MLR, RF, SVM

and LR models are not capable of handling the multi-dimensional data (formed as matrices), for these the output of the PAA was re-organised into a single-dimensional input by vectorising the matrix used as below:

$$Input = \begin{Bmatrix} x_{11} & x_{12} & ... & x_{sn} \end{Bmatrix}$$

The last pre-processing task before training the predictive models with the input data was data scaling. The experimental dataset was scaled using the normalisation technique that re-scales the ranges of each of the features to be between zero and one using minimum and maximum values of that feature.

Predictive Models, Experimental Setups, and Evaluation of Model Performance
As a baseline comparison, we employed the Multiple Logistic Regression (MLR) model used by Wells et al., and compared the results from this with those from four commonly used machine learning models.

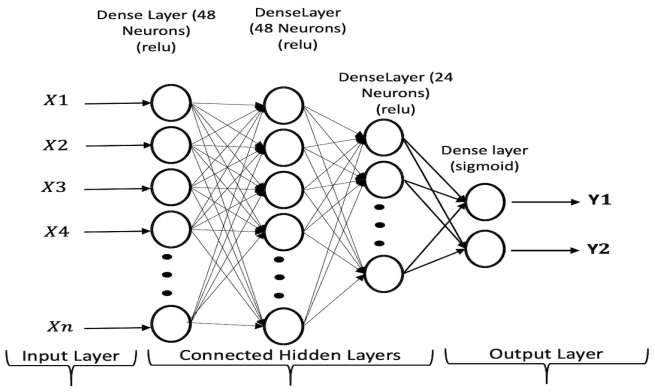
The MLR model is used to create a mathematical equation that can best calculate the probability of a value by the assigning weights (coefficients) to the independent variables (features) based on their importance (J. H. McDonald 2009). In this study we employed the same approach used by Wells et al. by which the continuous features were fitted into the MLR model using Restricted Cubic Splines (RCS) technique with 3-knots. When using the longitudinal input, the variables that caused collinearity were excluded.

The machine learning models used in this work are: Random Forest (RF), Logistic regression (LR) and Support Vector Machine (SVM)(introduced in section 3.2 of Chapter 3). Finally, the deep learning model used is the Multi-layer perceptron (MLP) (introduced in section 3.1 of Chapter 3).

Similar to the approach used in Chapter 6, the optimisation (selection of structures and hyperparameters) for the machine learning models (conventional and deep) applied in this chapter was performed by analysis of the empirical results obtained using the KAIMRC dataset. Fine-tuning of the structure and hyperparameters was performed for all models employed. For RF, the quality function used in the employed RF model is Gini, with a value of 100 for the number of trees parameters. The kernel function used in SVM model employed is Radial Base Function (RBF) with a value of 1 for the cost parameter (C).

The optimisation process for the deep learning model (MLP) used in this chapter involved tuning the neural network structure (e.g. number of hidden layers and neurons) and hyperparameters such as the activation functions (detailed in chapter 3), optimisers, and loss functions. The optimised structure of the deep learning model used in this study contained three hidden layers. The number of neurons in the hidden layers were 48, 48, and 24, respectively. The final layer (the output layer) contained 2 neurons for the final output of the model (Y1 for normal or Y2 for elevated HbA1c). A relu activation function was used in the three hidden layers and a sigmoid in the output layer. The detailed structure of the deep learning model is shown in Figure 7.13. The model was trained using an Adam optimiser with Mean Squared Error as the loss function.

Figure 7.13: The structure used for multi-layer perceptron (MLP) trained with the longitudinal data.



The optimised structure of the deep learning model used when using top available features

contained three hidden layers. The number of neurons in the hidden layers were 216, 216, and 108, respectively. The final layer (the output layer) contained 2 neurons for the final output of the model. A relu activation function was used in the two hidden layers and a sigmoid in the third and the output layers. The model was trained using an Adam optimiser with binary cross entropy as the loss function.

The models all employed the same data pre-processing, training, and testing techniques. The models were validated using the 10-fold cross-validation technique. For the deep learning model, 100 epochs were used to train each fold. We used the AUC-ROC, overall accuracy, and F1-score measures to evaluate and compare the performance of the models.

To uncover the importance the black box models (SVM and MLP) place upon each variable, we first compute the SHAP values and LIME scores for all samples in our dataset and then calculate the average absolute SHAP value and LIME score for each predictor.

7.3.2 Results

Since this section includes the investigation of employing advanced machine learning using two different feature sizes (the features used in the replication study (6) and the top available features (27)), the results for each one will be presented separately.

Results Using 6 Features

Table 7.9 shows the performance metrics obtained using the MLR, RF, SVM, LR, and MLP models with and without the longitudinal data using the 6 features. The results show that all models showed promising results for the task of predicting elevation levels of HbA1c using KAIMRC dataset. The RF, SVM, LR and MLP models trained without the longitudinal data achieved slightly better performance with regards to our main evaluation measure (AUC-ROC) than the statistical model employed by Wells et al.

Table 7.9 also shows that all models, including the MLR, achieved better performance using all reported measures when they are employed with the features from the patients' longitudinal

Table 7.9: Classifiers performance for current HbA1c levels prediction.

Model	**With longitudinal data**	**AUC-ROC, %(SD)**	**Accuracy, %(SD)**	**F1, %(SD)**
MLR	No	72.74% (4.15)	73.59% (3.79)	74.91% (5.12)
	Yes	73.49% (4.19)	74.30% (4.02)	75.11% (6.00)
RF	No	72.82% (1.18)	72.89% (1.04)	74.19% (1.08)
	Yes	74.02% (0.99)	74.06% (0.96)	75.15% (0.82)
SVM	No	73.69% (1.35)	73.88% (1.33)	75.76% (1.18)
	Yes	74.25% (1.11)	74.40% (1.08)	76.08% (0.92)
LR	No	73.18% (1.10)	73.17% (1.08)	73.96% (1.03)
	Yes	74.11% (1.15)	74.05% (1.13)	74.55% (0.98)
MLP	No	73.57% (1.21)	73.74% (1.20)	75.64% (1.16)
	Yes	74.51% (1.26)	74.66% (1.21)	76.27% (1.19)

data. The MLP with longitudinal data slightly outperformed all other models with respect to the AUC-ROC, overall accuracy, and F1-score.

Figure 7.14 summarises the 10-folds performance achieved for the set of measures where the models were trained without longitudinal data, and Figure 7.15 shows the performance where they were trained with the longitudinal data. Both figures show a more consistent prediction trend for RF, LR, SVM as well as MLP with and without longitudinal data, as the measures for these models show a small variation between the folds. As shown in Table 3, the SD values for MLR with and without longitudinal data are larger than for the rest of the models. This indicates that the machine learning models used can not only enhance the performance, but also improve the classification confidence for HbA1c prediction.

Table 7.10 shows the ranked order of importance of the set of predictors used for training the models. Further detail on the actual importance values for each model is provided in Appendix F (refer to Appendix G for more details of the LR and MLR calculator). Calculating the importance of the predictors for the MLR models using vectorised longitudinal data was not possible due to the collinearity caused by having multiple variables for BMI. The order of importance results obtained using the SHAP method for both the SVM and deep models are identical to those obtained using LIME, providing greater confidence in the explainability methods used.

Figure 7.14: Boxplot showing the details of the 10-folds performance of all models trained without longitudinal data.

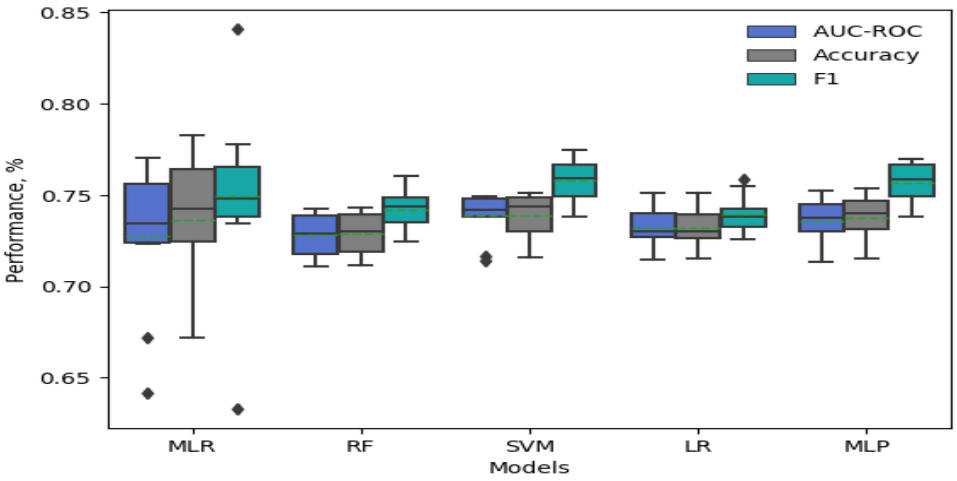


Figure 7.15: Boxplot showing the details of the 10-folds performance of all models trained with longitudinal data.

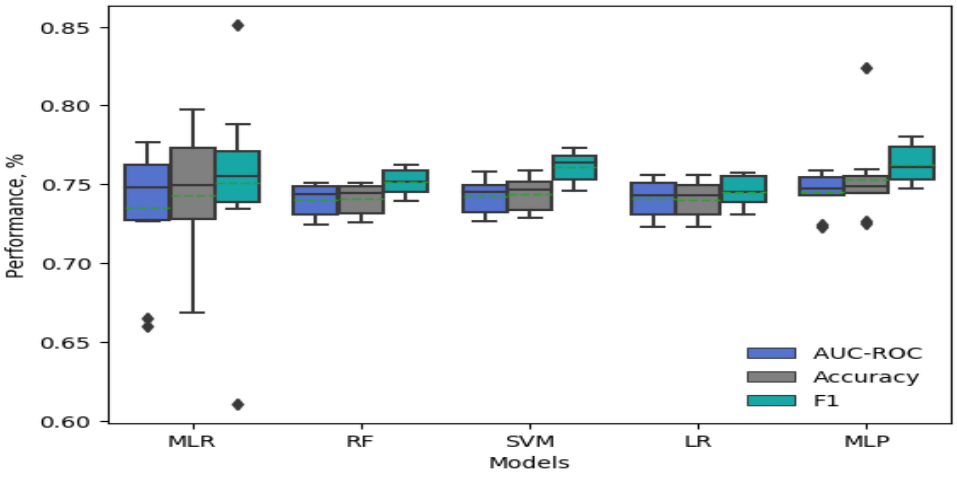
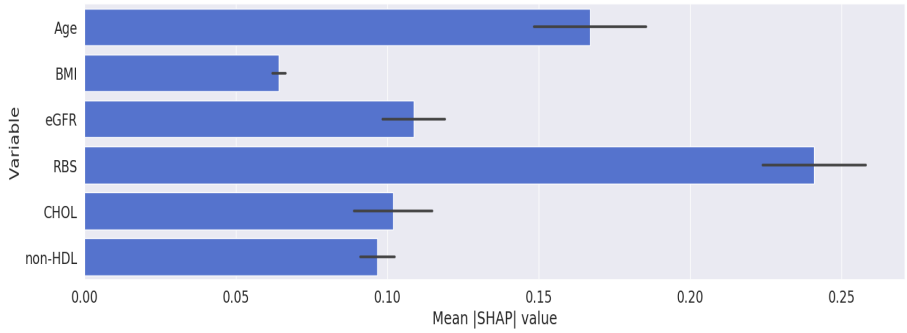


Table 7.10: Order of importance of predictors for the models used for HbA1c prediction.

Model	**With longitudinal data**	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**
MLR	No	Age	RBS	BMI	CHOL	Non-HDL	eGFR
RF	No	Age	RBS	BMI	eGFR	CHOL	Non-HDL
	Yes	RBS	Age	CHOL	eGFR	Non-HDL	BMI
SVM (SHAP & LIME)	No	Age	RBS	BMI	Non-HDL	CHOL	eGFR
	Yes	RBS	Age	CHOL	Non-HDL	BMI	eGFR
LR	No	RBS	Age	Non-HDL	CHOL	BMI	eGFR
	Yes	RBS	Age	Non-HDL	eGFR	CHOL	BMI
MLP (SHAP & LIME)	No	RBS	Age	Non-HDL	CHOL	BMI	eGFR
	Yes	RBS	Age	eGFR	CHOL	Non-HDL	BMI

Table 7.10 and the figures in Appendix F show that all of the models are heavily and interchangeably reliant on Age and RBS when making classification decisions. The RF and SVM models, when trained with longitudinal data, ranks RBS over Age. Interestingly, for the models trained with longitudinal data, BMI is ranked lower than when the models are trained without longitudinal data. However, the value produced by SHAP and LIME for the BMI variable from this model is still not insignificant (see Figure 7.16 and the figures in Appendix F). This indicates that models are able to find subtle relationships in the longitudinal data that are more relevant to the prediction than BMI, rendering it less important.

Figure 7.16: Relative importance of predictors obtained from MLP trained with longitudinal using SHAP.



When using the MLP and LR models trained on the longitudinal data the eGFR variable is ranked higher than CHOL and BMI, in contrast to when these are trained on the current visit only. None of the other models trained with the current visit only, except RF, consider it

important. Again, we ascribe this to the information that the model learns from the variations of eGFR values between a patient's visits (longitudinal EHR data).

SHAP values are calculated on the sample level. Figures 7.17 and 7.18 illustrate the SHAP values for two randomly selected samples from our dataset. These figures highlight how different inputs have different SHAP values. The patient in Figure 7.17 (for whom our model correctly predicts elevated HbA1c levels (>=5.7%)) has a higher RBS value than the patient in Figure 7.18 (for whom our model correctly predicts normal HbA1c levels (<5.7%)). This explains why our MLP model places much more importance on the RBS value of the patient in Figure 7.16.

Figure 7.17: An example shows the SHAP values for randomly selected sample with elevated HbA1c levels (>=5.7).

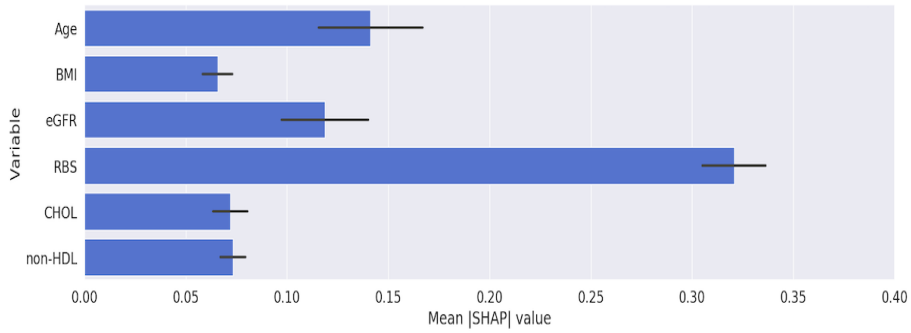
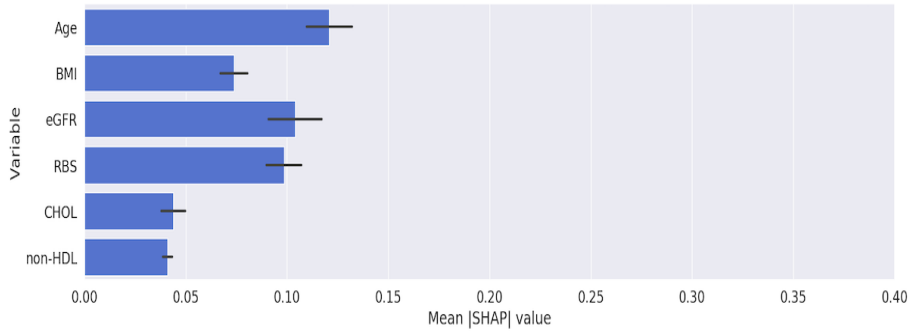


Figure 7.18: An example shows the SHAP values for randomly selected sample with normal HbA1c levels (<5.7).



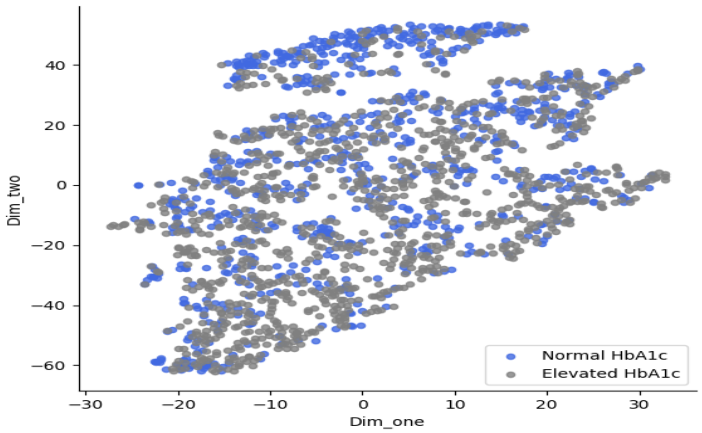
A statistical hypothesis test (Paired Student's t-test) (Urdan 2016) was used to evaluate the mean performance of the predictive models employed (without longitudinal data vs. with longitudinal data) using the cross-validation accuracy scores. A P value of (0.05) as a significance level was used to evaluate the statistical significance for the obtained AUC-ROC accuracy scores achieved by the models (P values are presented in Table 7.11). The statistical hypothesis test results in Table 7.11 show that there was a significant difference between the models (except for the MLR) when trained without vs. with longitudinal data. Therefore, the null hypothesis can be rejected for RF, SVM, LR and MLP models using the significance level of (0.05).

Table 7.11: Statistical hypothesis test results for the obtained AUC-ROC accuracy scores obtained by the models employed.

Model	AUC-ROC without longitudinal data	AUC-ROC with longitudinal data	P value
MLR	72.74%	73.49%	.1654
RF	72.82%	74.02%	.0003
SVM	73.69%	74.25%	.0075
LR	73.18%	74.11%	.0004
MLP	73.57%	74.51%	.0012

The task of predicting HbA1c elevation risk can be challenging. Figure 7.19 demonstrates this challenge by visualising the overlap in the test data between the two classes (pre-diabetic with >=5.7%) and (normal with <5.7%) using t-SNE (Maaten and Hinton 2008). We avoided intensive feature engineering techniques in the sampling approach used. However, the approaches adopted are able to achieve promising results with an accuracy of 74.51% using MLP with historical data.

Figure 7.19: Visualisation using t-SNE for randomly selected subset of the data.



In summary, the RF, SVM, LR and MLP models show better results for predicting the current HbA1c elevation levels using EHR data when compared to the model employed by Wells et al. (with increases of up to 1.77%). The results also emphasise that the HbA1c predictive models can exhibit more learnability when they are employed with the patient longitudinal observations that are normally available from EHR systems.

Results Using Top Available Features

Table 7.12 shows the performance metrics obtained using MLR, RF, SVM, LR and MLP models. All models including the MLR showed improvement on the prediction performance using all reported measures when employed with top variables that were available in the KAIMRC dataset with and without longitudinal data. The MLR model without longitudinal data achieved 75.58% for the AUC-ROC when more variables are included, compared to 72.74% using same population with only 6 variables.

Similarly, the prediction performance for RF, SVM, LR and MLP models without longitudinal data is also improved (RF from 72.82% to 76.21%, SVM from 73.69% to 75.89%, LR from 73.18%

Table 7.12: Classifiers performance for current HbA1c levels prediction with inclusion of top available variables.

Model	With longitudinal data	AUC-ROC, %(SD)	Accuracy, %(SD)	F1, %(SD)
MLR	No	75.81% (3.79)	76.42% (3.85)	77.04% (5.22)
	Yes	76.40% (3.45)	76.97% (3.53)	77.37% (5.17)
RF	No	76.21% (0.82)	76.26% (0.83)	77.29% (0.90)
	Yes	77.21% (0.46)	77.23% (0.45)	78.42% (0.50)
SVM	No	75.89% (1.26)	75.93% (1.27)	76.94% (1.44)
	Yes	76.81% (0.80)	76.84% (0.78)	77.82% (0.83)
LR	No	75.14% (1.08)	75.15% (1.06)	76.03% (1.12)
	Yes	76.17% (1.11)	76.17% (1.11)	76.92% (1.23)
MLP	No	76.46% (0.89)	76.46% (0.96)	77.14% (1.63)
	Yes	77.37% (0.69)	77.38% (0.58)	78.25% (0.62)

to 75.14%, and MLP from 73.57% to 76.46%). Using only the patient current visit data, the RF and MLP achieved the best performance with 76.21% and 76.46% accuracy for AUC-ROC.

All models showed better performance when trained with longitudinal data. The MLP and RF models outperformed all models when trained using longitudinal data with 77.37% and 77.21% accuracy, respectively, for the main measure used (AUC-ROC). Once again the SD values are greater for MLR models with and without longitudinal data than for the rest of the models. Figure 7.20 summarises the 10-fold measurements obtained for models trained without longitudinal data and Figure 7.21 for models trained with longitudinal data.

As shown in Table 7.12, the best performing models, RF and MLP with longitudinal data, have the smallest SD value. The results also show that the SD values for all predictive models when employed with longitudinal data tend to decrease for the measures used, except for LR.

Tables 7.13 and 7.14 show the order of importance of predictors for the models used in this study (the order of importance is presented in two tables for visualisation reason). The tables show that all of the models (MLR, RF, SVM, LR, and MLP) trained with and without longitudinal data agree on the high importance of Age and RBS variables. The models also show close level of agreement on the lowest ranked predictors such as Sodi, MPV, non_HDL, CI, and CHOL.

Figure 7.20: Boxplot showing the detailed performance of the models used with inclusion of top available variables and without longitudinal data for HbA1c elevation prediction.

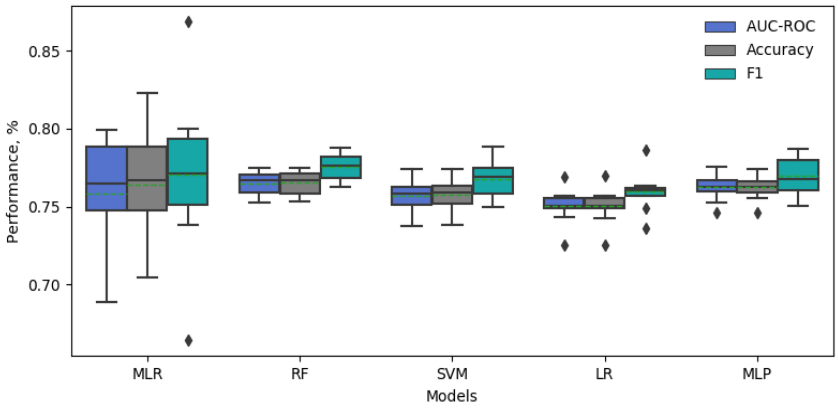


Figure 7.21: Boxplot showing the detailed performance of the models used with inclusion of top available variables and the longitudinal data for HbA1c elevation prediction.

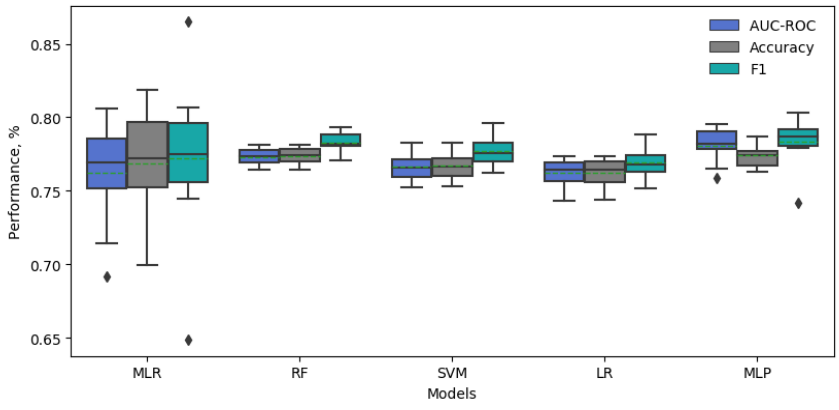


Table 7.13: Order of importance of predictors for the models used for HbA1c prediction.

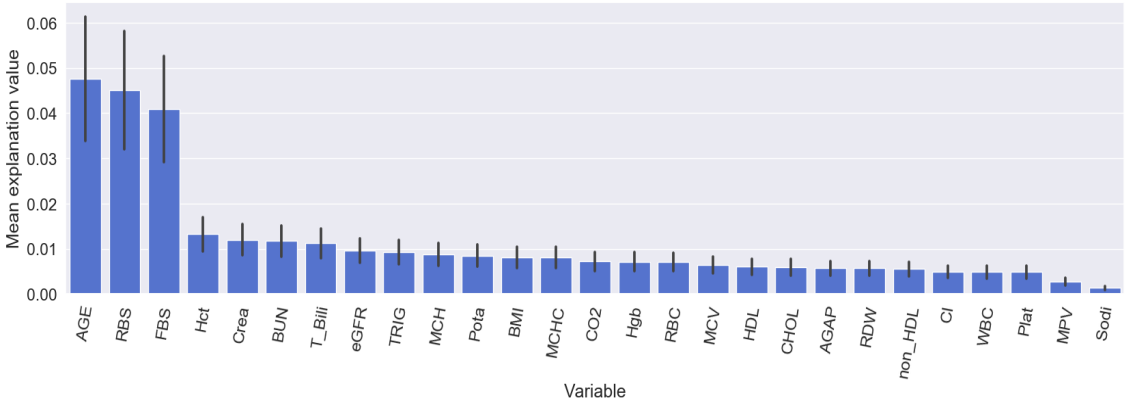
Model	Longitudinal data	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th	13th
MLR	No	Age	RBS	FBS	Crea	BMI	RDW	TRIG	T_Bili	Pota	MCHC	CI	WBC	CHOL
RF	No	Age	RBS	FBS	eGFR	TRIG	BMI	HDL	WBC	BUN	T_Bili	Crea	MCH	RDW
	Yes	Age	RBS	FBS	eGFR	TRIG	T_Bili	HDL	BUN	BMI	WBC	MCHC	MCH	RBC
SVM	No	Age	RBS	FBS	BMI	HDL	TRIG	BUN	Hct	MCV	RBC	Hgb	Pota	Crea
	Yes	RBS	Age	Hct	FBS	BMI	TRIG	Hgb	RBC	Pota	HDL	BUN	CO2	eGFR
LR	No	RBS	Age	FBS	BMI	Crea	TRIG	RBC	MCV	HDL	Hct	Hgb	eGFR	CO2
	Yes	RBS	Age	Hct	Crea	eGFR	FBS	Hgb	BMI	TRIG	BUN	Pota	RBC	CO2
MLP	No	Age	FBS	RBS	Crea	BUN	Hct	BMI	TRIG	T_Bili	MCV	RBC	Hgb	RDW
	Yes	Age	RBS	FBS	Hct	Crea	BUN	T_Bili	eGFR	TRIG	MCH	Pota	BMI	MCHC

Table 7.14: Order of importance of predictors for the models used for HbA1c prediction (cont).

Model	14th	15th	16th	17th	18th	19th	20th	21st	22nd	23th	24th	25th	26th	27th
MLR	BUN	MCV	Sodi	non_HDL	Hgb	AGAP	MCH	eGFR	Hct	Plat	CO2	MPV	RBC	-
RF	MCV	Plat	CHOL	non_HDL	MCHC	RBC	MPV	Pota	Hct	Hgb	AGAP	CI	CO2	Sodi
	Crea	RDW	MCV	Plat	Pota	MPV	AGAP	non_HDL	CHOL	CI	Hct	Hgb	CO2	Sodi
SVM	AGAP	RDW	MCH	CI	MCHC	Plat	CO2	T_Bili	WBC	eGFR	CHOL	MPV	non_HDL	Sodi
	MCV	MCHC	RDW	AGAP	Crea	MCH	CI	Plat	non_HDL	CHOL	T_Bili	MPV	WBC	Sodi
LR	Pota	AGAP	RDW	MCH	Plat	BUN	T_Bili	CI	MCHC	WBC	non_HDL	MPV	Sodi	CHOL
	MCV	T_Bili	MCH	HDL	AGAP	RDW	Plat	CI	non_HDL	MCHC	CHOL	MPV	WBC	Sodi
MLP	Pota	HDL	AGAP	MCH	MCHC	CI	CO2	eGFR	WBC	CHOL	non_HDL	Plat	MPV	Sodi
	CO2	Hgb	RBC	MCV	HDL	CHOL	AGAP	RDW	non_HDL	CI	WBC	Plat	MPV	Sodi

All models trained with and without longitudinal data are heavily reliant on Age and RBS (and FBS for RF and MLP models specifically). Refer to Appendix I for more details about the relative importance of predictors for all models. Figure 7.22 visualises the order of importance for the best performing model, MLP with longitudinal data.

Figure 7.22: Order of importance of predictors for the MLP model trained with longitudinal data using SHAP.



The use of longitudinal data for the RF does not have major effect on the order of importance of the highest ranked predictors, except Total Bilirubin (T_Bili) which is ranked higher and BMI which is ranked lower when the model trained with longitudinal data. In fact, the models, SVM, LR and MLP, rank BMI lower than when trained without longitudinal data.

The Haematocrit (Hct) is not of much importance for the RF with and without longitudinal data. We see similar behaviour for this variable on the LR trained without longitudinal data. However, the Hct shows more importance for LR when trained with longitudinal data, SVM and MLP when trained with and without the historical data.

The eGFR shows higher level of importance for the SVM, LR, and MLP models when the models trained with longitudinal data. When trained without longitudinal data, it is ranked 23rd for SVM , 12th for LR and 21st for MLP. However, when these models trained with longitudinal data the eGFR is ranked 13th for SVM, 5th for LR, and 8th for MLP. Also, LR and MLP rank Creatinine (Crea) higher than the SVM and RF.

The tables also show that the TRIG and HDL from the lipid profile laboratory tests (the lipid profile tests includes: TRIG, CHOL, non-HDL, HDL) are ranked with higher importance when compared to CHOL and non-HDL. The ranking order presented in Tables 7.13 and 7.14 shows that the HDL is always ranked lower when the same model trained with longitudinal data, except in RF model. However, the importance of HDL is still not insignificant in all models.

The order of importance presented in Tables 7.13 and 7.14 confirms the results obtained in the section 7.3. This work shows that the rankings of the variables are different for the models employed for the prediction of HbA1c using KAIMRC dataset. It also shows that the use of longitudinal data effects the relative importance order. For instance, the eGFR always gain higher importance for all models when trained with longitudinal data, except RF.

7.3.3 Discussion and conclusion

EHR systems were adopted for the purpose of improving healthcare outcomes and were not originally intended for research purposes (Stanley Xu et al. 2014). Patient data stored in EHR

systems can be irregular, as lab instructions are carried out with different frequencies based on the physician's decisions and a patient's visit patterns. It is very common that medical data extracted from EHR systems suffer from problems such as irregularity, incompleteness, and noisy and imbalanced data (Miotto et al. 2016). These can be challenging obstacles for any technology used for predictive analytics.

Remembering that the data available in the EHR systems are not targeted to be used to solve a specific research problem or build predictive models, the availability of the variables (laboratory tests) is mainly dependent on the clinical procedures and physicians decisions. Cost and geographical factors can also affect the availability of a particular tests. For instance, for patients in Saudi Arabia, it is recommended that they are regularly and intensively screened for hyperglycemia, whereas this is not the case in other regions in the world (Al-Zahrani et al. 2019).

The sampling approach used did not affect the balanced nature of the dataset used. As shown in Figure 7.10, there were 56,185 unique patients before removing the records with one or more missing values. The number of unique patients with elevated HbA1c levels (>=5.7) before removing the incomplete records was 27,354 with 48.68% (27,354/56,185). The number of unique patients with normal HbA1c levels was 28,831 with 51.32% (28,831/56,185). We would argue that the absence or the presence of the HbA1c readings is not random. Being a sample collected from the population of Saudi Arabia, the likelihood of a patient taking an HbA1c test is large because of the prevalence of diabetes (Al-Zahrani et al. 2019). This may affect the reproducibility of this work using different populations from different countries especially those with lower rates of diabetes.

Our dataset contained only three years of patient data, which limits the number of patient visits recorded. Figure 7.23 shows the number of visits made by patients from 2016 to 2018 over age groups. This shows that the majority of the patients have made relatively few visits. The trend in the number of visits made by patients during this period is consistent across all age groups. Figure 7.24 details the number of visits made by patients after removing the outliers (patients making more than 14 visits). 52% (8713/ 16818) of the patients have made four visits or fewer

during the three years (1.3 visit per year). This also justifies the size of the sliding window ($s = 3$) as the optimal input size for the deep learning model. However, we hypothesise that the longitudinal behaviour of the features used can be enriched by employing more values obtained over longer periods. Therefore, incorporating more features and their longitudinal behaviour over longer periods into the models used in this study would be likely to improve the model's prediction performance.

Figure 7.23: Boxplot showing the trend in number of visits made by patient over age groups.

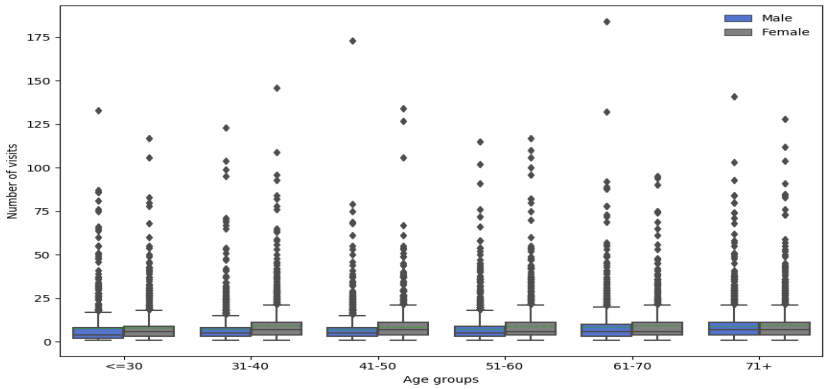
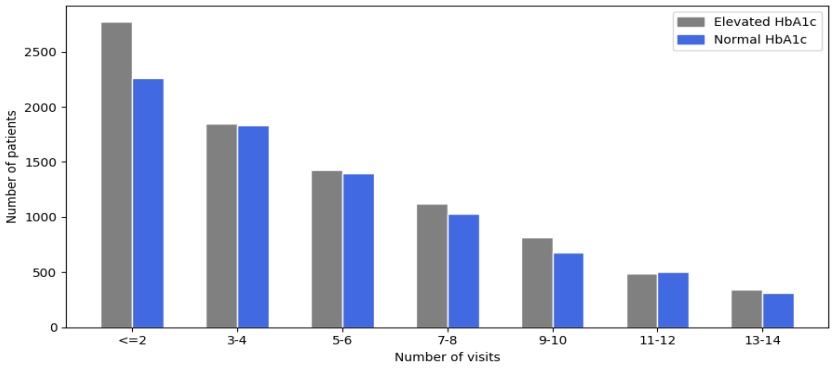


Figure 7.24: The details for the number of visits made over number of patients.



Variations in the data/model produce slightly different attribution values. However, due to the critical nature of many healthcare applications, it is always important to verify that our models make 'sensible' predictions. Without the use of SHAP/LIME, this would be hard to verify for any non-linear model. Although it is possible to see that the models have high performance, we would be unable to verify that the model is not making spurious correlations. Furthermore, through the use of SHAP, we can verify that MLPs trained on the longitudinal data are learning to use the extra information contained in the longitudinal data (as indicated by the higher importance of eGFR), allowing us to pinpoint the reason these models gain higher performance.

In this section, we used the common approach for calculating the importance of variables for RF models using the Gini impurity technique (Nembrini, König and M. N. Wright 2018). In general, the impurity importance is calculated by summing the impurity decrease/reduction of the nodes in the a split of the tree. However, we have also used SHAP in explaining the RF model. The result obtained using SHAP showed insignificant differences with those obtained from using Gini impurity technique (refer to Appendix I for more details about the order of importance obtained for RF using SHAP). This can add more robustness to the SHAP results obtained for the blackbox models, SVM and MLP.

With regards to studying the effect of temporal dependencies in the employed variables in data subset (C) and (D), we also involved applying LSTM, BiLSTM and GRU models. Table 7.15 reports the results obtained employing these models with longitudinal data. The LSTM, BiLSTM and GRU models showed promising results. However, there was no improvement on the performance when employing these models. This suggests that directly modelling the temporal dynamics in the data is not very helpful. This could be due to the short lengths of the time series, or to weak temporal dependency.

It can be difficult to have high number of variables included in the medical predictive models (Rajkomar et al. 2018). The presence of variables with a large number of missing values may increase the pre-processing effort or decrease the sample size used for training the models, depending on the sampling method adopted. The more variables are used, the less complete

Table 7.15: LSTM, BiLSTM and GRU classifiers performance with longitudinal data.

Model	**Data subset**	**AUC-ROC, %(SD)**	**Accuracy, %(SD)**	**F1, %(SD)**
LSTM	C	74.18% (1.33)	74.31% (1.23)	76.00% (1.04)
	D	76.87% (0.85)	76.97% (0.80)	77.88% (1.58)
BiLSTM	C	74.09% (1.23)	74.17% (1.15)	75.62% (0.90)
	D	75.94% (1.55)	76.08% (1.44)	77.48% (1.43)
GRU	C	74.35% (1.50)	74.46% (1.39)	75.98% (1.06)
	D	77.13% (0.96)	77.12% (0.99)	77.77% (1.58)

records are anticipated from EHR datasets. However, this work has also investigated including more than 20 variables. However, those variables did not help to improve the performance of the models. This can be ascribed to the large number of missing values in these variables or low relatedness to the HbA1c when other important variables (i.e. age and RBS) exist.

As aforementioned, the data stored in EHR systems is normally available beforehand for use by predictive models. To benefit from this, we aimed to employ as many as possible predictors (features) in our models. Specifically, the predictors that did not heavily effect the sample size of the data subset employed (the top available variables that have fewer missing values). Although, one of the main objective for this approach is to find/discover hidden correlation between the target variable (elevated levels of HbA1c) and those predictors, however, finding the predictors/variables that showed lower (or no) correlation/importance in the population employed can also be of interest in this area of research.

We hope that this work will encourage further investigation into the outcomes of this work from the clinical perspectives. Most importantly, studying the association between the variables used on the HbA1c levels for patients with no history of hyperglycemia (patients with pre-diabetes). The clinical investigation may include studying the association/importance of the clinical variables for patients with diabetes vs. pre-diabetes. For example, there are many studies that have demonstrated the relationship between diabetes prevalence and BMI (Boffetta et al. 2011). However, a recent study by Rahmanian et al. (2016) showed that the BMI shows low association with pre-diabetic men patients (which can also confirm the results obtained in Tables 7.13 and 7.14).

As the Col-DAE (applied in Chapter 6) is based on the autoencoder technique, the performance of using small number of features that were originally involved in studies limited generalising the Col-DAE in this chapter (generalising this model by using more features and larger datasets is discussed in Chapter 9 as part of the future work). Similarly, generalising our findings to other datasets is difficult because of the accessibility and privacy restrictions that apply to medical datasets. For this reason, and because of the lack of similar studies that have employed machine learning for HbA1c prediction using EHR data, comparing the performance achieved by the models outlined in this work with those developed by other researchers will require the availability of alternative anonymised datasets.

To the best of our knowledge, this work is the first to investigate the performance of machine learning models used for predicting current HbA1c elevation risk for non-diabetic patients. It is also the first to investigate employing the longitudinal data that are normally stored on EHR systems to enhance the prediction of HbA1c elevation levels. Our findings show that all models employed achieve better results when a patient's longitudinal data are combined with current visit data, and the use of longitudinal data also affects the relative importance for the predictors used. Also, this work showed that the elevation levels of HbA1c test can be reliably predicted, with minimal effort in data preparation, from the data normally (top) available in the EHR systems

Epilogue

This chapter provided an investigation about the prediction of current HbA1c for non-diabetic patients. The chapter started with replicating a replication study to validate, evaluate, and identify the strengths and weaknesses of the use predictive models for HbA1c prediction using KAIMRC population. This chapter also investigated the use of longitudinal data in EHR systems with advanced machine learning approaches in HbA1c predictions. The following chapter will investigate the second challenge of this thesis; mortality risk predictions using machine learning with imbalanced EHR data

Chapter 8

Stacked Denoising Autoencoders for Mortality Risk Prediction Using Imbalanced Clinical Data

Prologue

Quick and accurate prediction of mortality can be critical for physicians to make intervention decisions. In this chapter, we introduce a predictive deep learning model aiming to evaluate the mortality risk of in-hospital patients. Stacked Denoising Autoencoder (SDA) is trained using the KAIMRC dataset which is naturally imbalanced with regards to in-hospital mortality rates. The results are compared to common deep learning approaches using different methods for data balancing.

This chapter sought to address the research question below (mentioned in the Introduction chapter):

- - Can the risk of in-hospital patient mortality be measured accurately by machine learning models using EHR data?

Declaration: Description of this study as presented in this chapter is largely as published on the following publication: Alhassan Z, Budgen D, Alshammari R, Daghstani T, McGough AS, Al Moubayed N. *Stacked denoising autoencoders for mortality risk prediction using imbalanced clinical data.* In2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) 2018 Dec 17 (pp. 541-546). IEEE (Alhassan, Budgen, Alshammari, Daghstani et al. 2018).

The references and notations have been altered, cross-references have been added and some stylistic changes have been made for the consistency throughout this thesis.

8.1 Introduction

Predicting the mortality risk of patients is a major concern for physicians in the medical domain. Accurate prediction of mortality (referred in some studies as discharge type) can introduce improved healthcare services to aid with the survival of patients. The quick and timely interpretation of clinical data is needed by physicians to improve patient outcomes (Luo et al. 2016). Thus, the early prediction of in-hospital mortality risk is a major area of interest for research.

In this chapter, we investigate the performance of Autoencoder models to predict the mortality risk of in-hospital patients using the KAIMRC clinical data. Since the proposed data is naturally imbalanced, we formulate the mortality risk classification problem as a problem of anomaly detection (patient discharged home coded as normal, and died as abnormal) and use the Autoencoders with unlabelled data.

To the best of our knowledge, this work is the first (at time of publishing) to investigate the use of a predictive model for mortality prediction in general (regardless of the health problem)

using the Stacked Denoising Autoencoder (SDA). The model is trained using a single class of the KAIMRC dataset. The main contributions of this chapter are:

- Adopting the autoencoders machine learning technique (stacked denoising autoencoder) for predicting the mortality risk after 24 hours of in-hospital patients admission using EHR clinical data.
- Uniquely tackle the imbalance data problem that naturally exists in the mortality datasets. This include comparing the results obtained using our approach with common approaches used to handle the data imbalance problem.

8.2 Methodology for the Mortality Risk Prediction

We follow a unique approach to investigate the predictability of in-hospital mortality risk using EHR data. Normally, the in-hospital patient mortality after admission is rare compared with patients being discharged home. The distribution of the dataset classes can affect the performance of most classification algorithms. Thus, depending on the dataset and it's size, this problem is traditionally solved either by artificially creating more samples for the minority class (over sampling) or eliminating samples from the majority class (under sampling). To increase the reliability of the outcomes of this work, we use real data only for testing the models. Unlike the approaches used in the related work that test the predictive models using mix of real and artificially generated data, this work investigate testing the models using real data only.

We propose the use of the SDA for mortality risk prediction. Unlike previous studies that formulated the mortality prediction into binary classification problem for specific health conditions and with the data of 48 hours after admission, this work investigates the performance of predicting an in-hospital patient's mortality risk in general(regardless of the health condition types) and after only 24 hours of patient admission using a unique dataset (KAIMRC). We compare our results against three commonly used base-line models: Support Vector Machine (SVM), Multi-Layer

Perceptron (MLP) and Long Short-Term Memory (LSTM) (LeCun, Bengio and Hinton 2015; Hochreiter and Schmidhuber 1997).

Figure 8.1 details the approach used to achieve the objectives of this work. The first task is preparing the data for the models. Then two methods of data scaling are used prior to build the models. To avoid training the comparative models using imbalanced data, an over-sampling method is used to generate more samples from the minority class data. Borderline Synthetic Minority Over-Sampling Technique (bSMOTE) and Support Vector Machine Synthetic Minority Over-Sampling Technique (SVM-SMOTE) over-sampling methods are used for balancing the data (Han, W.-Y. Wang and Mao 2005; N. V. Chawla et al. 2002). Those models are then trained using real data from the majority class and the artificially generated data from the minority class and tested using real data only from both classes. We also investigate the performance of these models without data oversampling.

For the Autoencoder models, AE and SDA, the models are trained using single-class only (the majority class) and tested using mixture of the real data from both the majority and minority classes.

8.2.1 Data Preparation

For the purpose of our analyses (predicting the mortality after 24 hours), only the data of the patients with two or more days of length of stay were considered. Patient visits with less than two days in hospital have been excluded. The patients discharged for administrative reasons have also been excluded, making the experimental data, KAIMRC data subset(E), subset size to be 3,557 patient visits (Table 8.1).

There are 86 features originally extracted: gender, age, service, specialty, visit type and 81 vital signs and laboratory results. Refer to Appendix B for more details (ICD10 code and description) about the variables employed in this data subset. To give an early prediction for the discharge type of a patient visit, the values for the first day of the patient visit are selected to train and

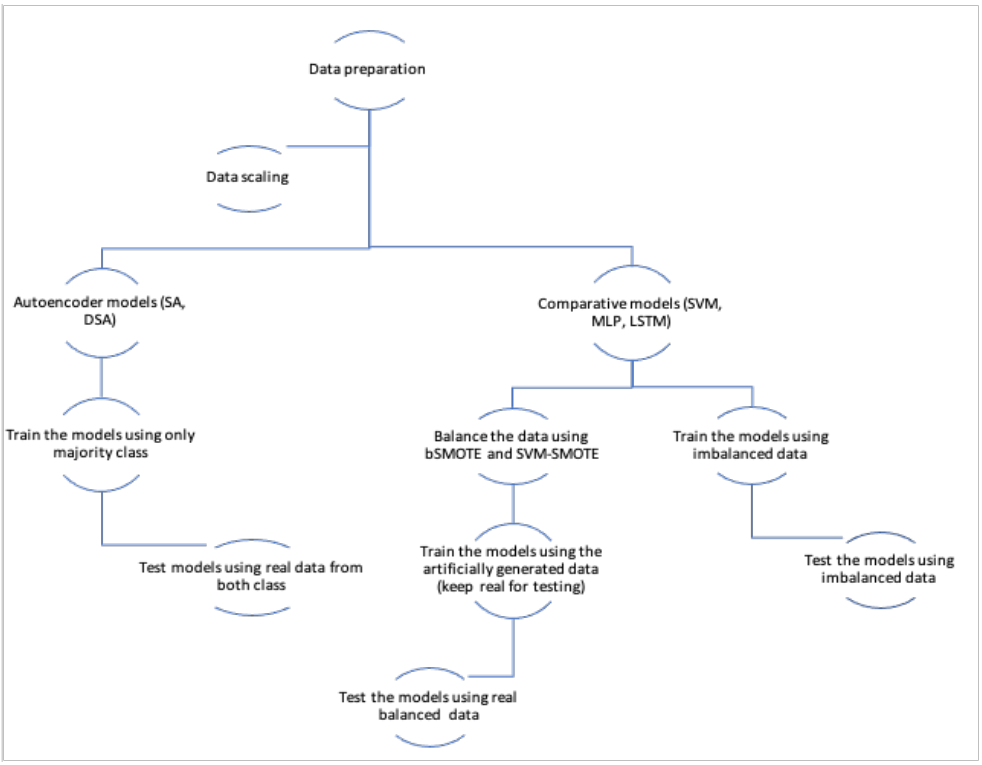


Figure 8.1: The approach used to investigate the prediction of in-hospital mortality risk.

test the the models. Some features are changing frequently as they may have been collected on an hourly basis, such as vital signs. In these cases, the average value for these readings on that day are used. In case of missing readings, the first available value for that readings taken on the next days is considered. If there are no readings taken in the whole visit, we consider it to be missing data and replace with zeros.

Figure 8.2 presents more details about the number of visits made by patients over gender distribution. The experimental data subset shows a slightly balanced distribution for patient visits over gender.

Table 8.1: Profile for the experimental dataset (KAIMRC data subset(E)).

Characteristic	**Overall**	**Used**
Number of patient visits	14,609	3,557
Number of features	500+	86
Number of different health conditions	99	97
Number of patient visit types	4	2
Number of discharge types	8	2

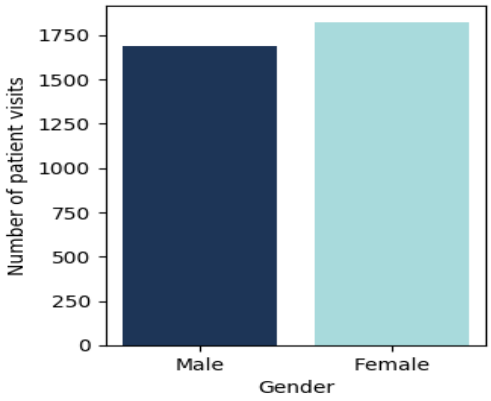


Figure 8.2: Patient visits over gender distribution (KAIMRC data subset(E)).

8.2.2 Interpretations Approach for Data Imbalance

The majority (95%) of the patient visits of KAIMRC dataset are labelled with "discharge home". The remaining 5% of the data are labelled with "patient died". As illustrated in Figure 8.3, the experimental dataset is therefore severely imbalanced. We propose investigating common solutions to overcome the problem of imbalanced data for the base-line models. Data over-sampling is applied on the samples with minority labels. The Synthetic Minority Over-Sampling Technique (SMOTE) (N. V. Chawla et al. 2002) is one of the common methods for data over-sampling. SMOTE is based on the synthetic creation of new examples of the minority class. The nearest neighbours of the created samples are randomly chosen, based on the number of needed examples.

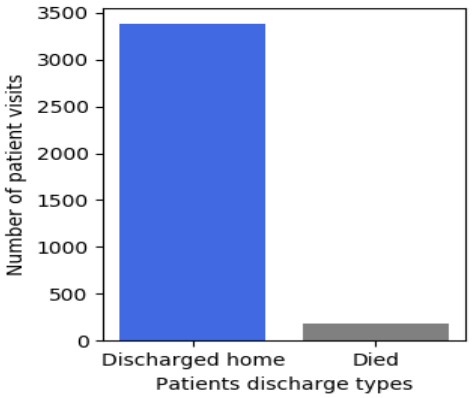


Figure 8.3: Patients discharge types (Discharge vs in-hospital mortality) distribution (KAIMRC data subset(E)).

There are several versions of the SMOTE algorithm used to evaluate and solve the imbalance problem. bSMOTE is a version of SMOTE that considers the examples close to the minority borderlines during the process of over-sampling (Han, W.-Y. Wang and Mao 2005; Nguyen, Cooper and Kamei 2011). Support Vector Machine SMOTE (SVM-SMOTE) (N. V. Chawla et al. 2002) is another version that uses SVM classifier to create number of synthetic examples around the negative class points (Y. Tang et al. 2009). In this work, we applied bSMOTE and SVM-SMOTE methods to overcome the problem of imbalanced data when using the supervised learning algorithms in the base-line models.

8.2.3 Data Scaling Methods

Data scaling is a process of making the ranges for the dataset features into the same scale (Ali et al. 2014). This process is usually part of the data pre-processing task. Normalisation and standardisation are two common methods for data scaling (Cao, Stojkovic and Obradovic 2016). Normalisation uses minimum and maximum values for each feature to re-scale values between 0 and 1. Standardisation changes the distribution of the feature values to be centred on 0 and the standard deviation of 1. These methods can be used as part of data pre-processing and before

building and feeding the models with the input data.

8.2.4 SDA Model and Experimental Setup

We investigate the use of Stacked Denoising Autoencoder (SDA) for predicting the in-hospital mortality risk for the patients KAIMRC dataset. As discussed in Chapter 3, the SDA is an extended version of the Autoencoder (explained in Chapter 3). SDA tends to force the hidden units to extract features from a corrupted version $\tilde{x}$ of the original input x. The input x for the DAE model is a sequences of patient observations as input (Eq:8.1).

$$x : x_1, x_2, ...x_n \tag{8.1}$$

The input will be decoded to help the network learn extracting important features from the patient visits data. Also the decoding will help in avoiding the identity function problem by undoing the corruption. The DAE tries to minimise the reconstruction error using the corrupted version of the input $\tilde{x}$ (Eq. 8.2).

$$L(x, g(f(\tilde{x}))) \tag{8.2}$$

Finally, we use the Mean Squared Error (MSE) function (Eq. 3.15 provided in Chapter 3) to calculate the reconstruction error after fitting the model with the test data.

Similar to the approach used in Chapters 6 and 7, the optimisation (selection of structures and hyperparameters) for the machine learning models (conventional and deep) applied in this chapter was performed by analysis of the empirical results obtained using the KAIMRC dataset. This involved tuning the neural network structure (e.g. number of hidden layers and neurons) and hyperparameters such as the activation functions (detailed in chapter 3), optimisers, and loss functions for the models used (MLP, LSTM, and autoencoders).

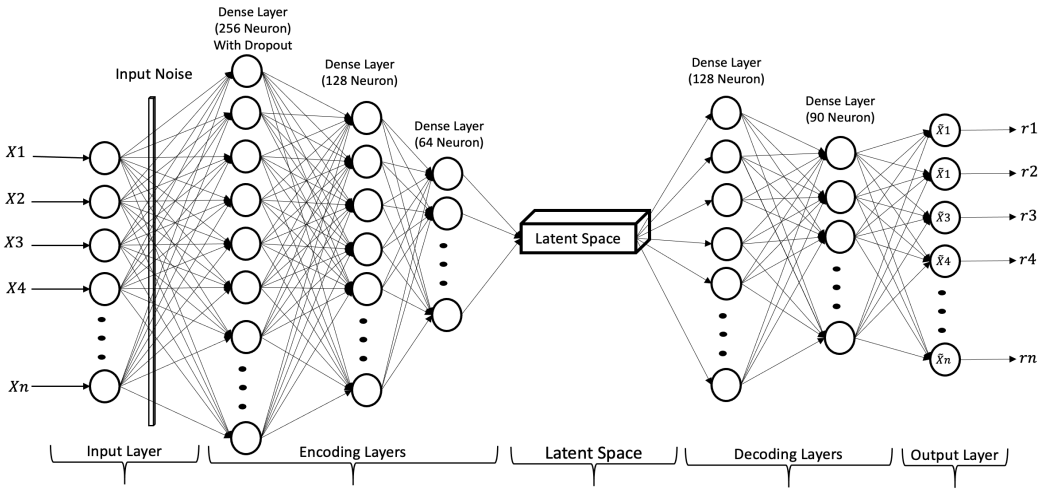


Figure 8.4: The structure for the stacked denoising autoencoder (SDA) model used.

The input layer for the optimised SDA neural network model contains three stacked hidden layers for the encoder, as shown in Fig 8.4. Prior to the encoder, a noise is added to the input layer using the Gaussian noise method (P. Vincent et al. 2010). The first layer of the encoder was attached with a dropout of 0.1. The number of neurons for the encoder layers are 256, 128 and 64 respectively. A tanh activation function is used in the first layer and relu is used for remaining encoder layers. The model also contains two hidden layers for the decoder. The first decoding hidden layer holds 128 neurons with relu and the second hidden layer holds 90 neurons with a tanh activation function.

The SDA model is trained and validated using 80% and 10% respectively of the majority class data only. The remaining 10% of the majority class is then used along with minority class data (anomaly samples) for testing. The model uses RMSprop optimiser with Mean Squared Error as loss function. Before preforming the prediction on the test data, the model was trained for 100 epochs. In our analyses, we investigated the performance of the model with the first 24 hours of patient data after admission. Since the test data are imbalanced, we report F1-macro, Recall-macro and Precision-macro scores to evaluate the performance of the models. A threshold

is then chosen to decide on the outliers (anomalies) based on the calculated MSE values (Eq. 3.15).

8.3 Results

Table 8.2 shows the performance metrics obtained using the SVM, MLP, LSTM, SA and SDA models. The models are trained with and without over-sampling. In case of over-sampling, bSMOTE and SVM-SMOTE are used for data balancing and training the models. To evaluate the models accurately and robustly, only real data from both classes was used for testing (no artificially generated samples were used for testing). For training the base-line models, real data from the majority class with the over-sampled data from the minority class were used. Table 8.2 shows the results when the models were trained with scaled data using normalisation and standardisation techniques. SDA model with normalisation, achieved an accuracy of 74.05% for F1-macro, 77.13% for Recall-macro and 72.92% for Precision-macro.

In Table 8.2, the results emphasise the impact of the method used for scaling the data when applying SA and SDA. Scaling the data using the normalisation in the SDA model achieved 74.05% for F1-macro while standardisation achieved significantly fewer results with 61.92% in the same model using the same data. The test data points with reconstruction error values are presented in Figure 8.5 and Figure 8.6.

The predicted data using standardisation contrasts with the reconstruction error value for normal and anomalous data points. The results also show that the Autoencoders (SA and SDA) can perform better when noise is added to the input layer. The Gaussian noise provides the Autoencoder with more generalised input which helps the model to detect the anomalies.

Table 8.2 also shows that balancing the data using oversampling algorithms (bSMOTE and SVM-SMOTE) for deep learning models (MLP and LSTM), only has a minor impact on the result compared to the imbalanced data. This is not the case for SVM models, which show better accuracy when using artificially balanced data.

Table 8.2: Classifiers performance for mortality risk prediction.

Model	**Over-Sampling**	**Scaling**	**F1 macro**	**Recall macro**	**Precision macro**
SVM	None	Norm	0.4845	0.5000	0.4700
		Stand	0.4868	0.4998	0.4745
	bSMOTE[a]	Norm	0.6195	0.7525	0.5968
		Stand	0.6416	0.7688	0.6106
	SVM-SMOTE[b]	Norm	0.6566	0.6960	0.6420
		Stand	0.6572	0.7652	0.6236
MLP[c]	None	Norm	0.6706	0.6581	0.7270
		Stand	0.6297	0.6291	0.6394
	bSMOTE	Norm	0.6442	0.6778	0.6270
		Stand	0.6535	0.6580	0.6521
	SVM-SMOTE	Norm	0.6502	0.6690	0.6394
		Stand	0.6384	0.6320	0.6502
LSTM[d]	None	Norm	0.6518	0.6319	0.6982
		Stand	0.6552	0.6405	0.6757
	bSMOTE	Norm	0.6628	0.6814	0.6539
		Stand	0.6483	0.6417	0.6616
	SVM-SMOTE	Norm	0.6598	0.6908	0.6562
		Stand	0.6394	0.6222	0.6667
SA[e]	None	Norm	0.7310	0.7403	0.7242
		Stand	0.6204	0.6130	0.6376
SDA[f]	None	Norm	**0.7405**	**0.7713**	**0.7292**
		Stand	0.6192	0.6120	0.6357

- [a]bSMOTE: Borderline Synthetic Minority Over-Sampling Technique.
- [b]SVM-SMOTE: Support Vector Machine SMOTE.
- [c]MLP: Multi-layer perceptron.
- [d]LSTM: Long-Short Term Memory.
- [e]SA: Stacked Autoencoder.
- [f]SDA:Stacked Denoising Autoencoder.

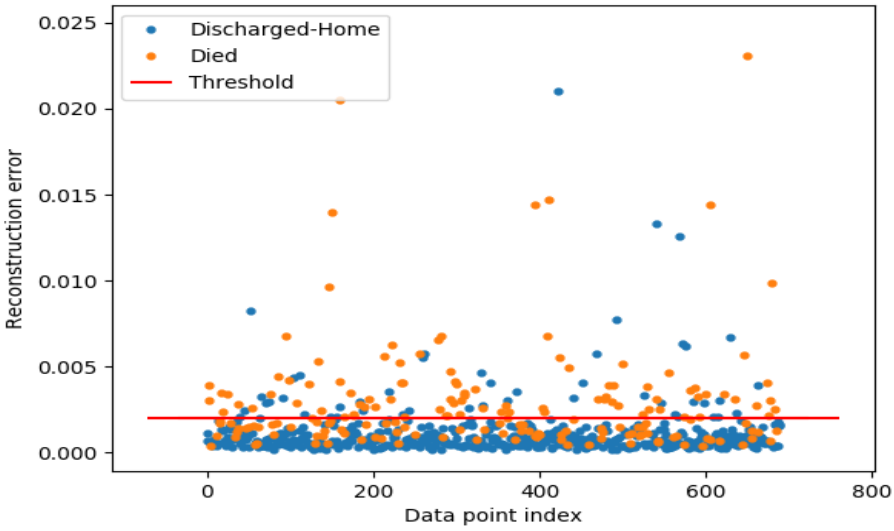


Figure 8.5: SDA predicted data points space using normalised data.

8.4 Discussion

Clinical data, such as evaluations, treatments, vital sign and lab test results, are usually observed and recorded in hospital systems. Making use of such data to help physicians to evaluate the mortality risk of in-hospital patients provides an invaluable source of information that can ultimately help with improving healthcare services. In particular, quick and accurate predictions of mortality can be valuable for physicians who are making decisions about interventions.

The task of predicting the mortality risk can be challenging. Fig 8.7 demonstrates this challenge by visualising the overlap in the SDA latent space in the test data between the two classes using t-SNE (Maaten and Hinton 2008).

In this chapter, we investigated a novel application of the Stacked Denoising Autoencoder (SDA) for in-hospital patients mortality risk prediction. The model, using patients clinical data from a variety of health conditions and without intensive feature engineering, achieved promising results using only the first 24 hours of data after patient admission.

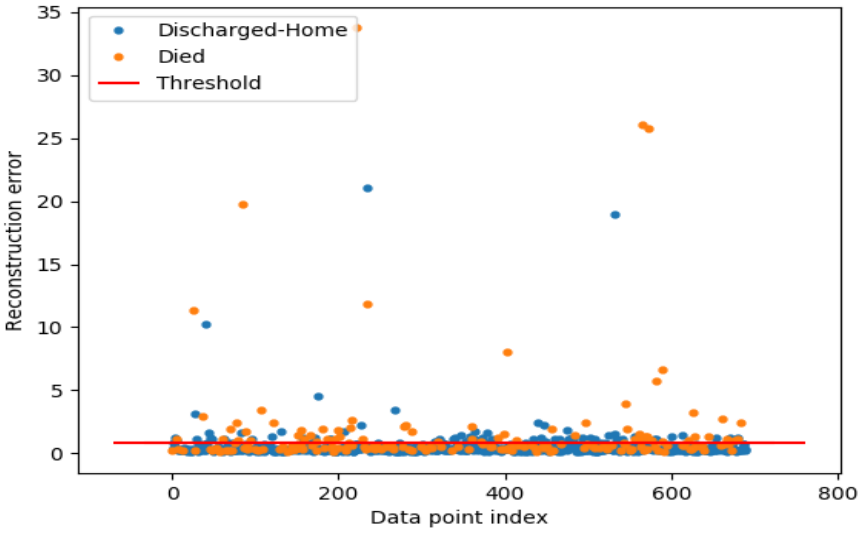


Figure 8.6: SDA predicted data points space using standardised data.

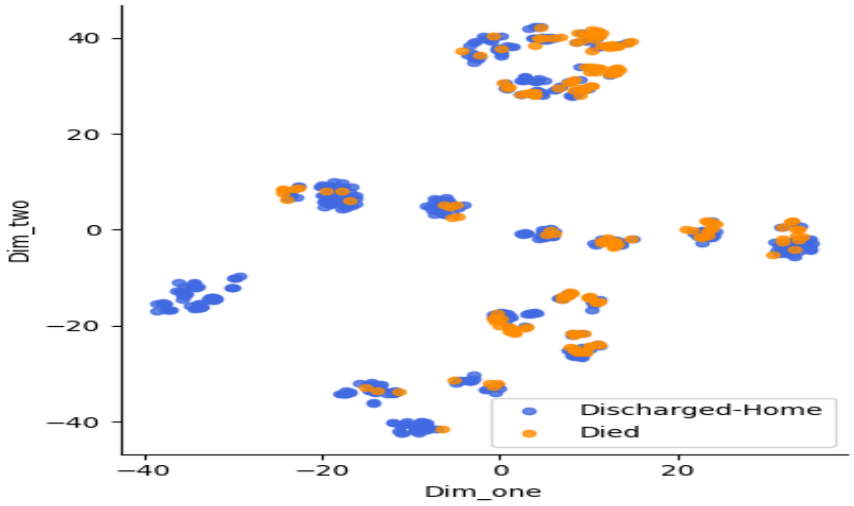


Figure 8.7: SDA latent space visualisation for test data using t-SNE.

The model was trained using only the majority class data of the KAIMRC dataset. It was tested using a mixture of majority and minority classes. In proposed framework only real data from both classes was used for testing. This evaluation approach provided more robustness to the results when compared to the models in the related work that use over-sampled data for training and also testing thier models.

Our model outperformed the base-line classifiers and achieved an accuracy of 74.05% for F1-macro, 77.13% for Recall-macro and 72.92% for Precision-macro. The SDA model gives promising mortality risk prediction results within only 24 hours from patient admission. This can be very significant for clinicians to make quicker intervention decisions to provide an improved healthcare services to the patients (especially for those under the risk of mortality). The results are compared to those from common deep learning approaches, using different methods for data balancing. The employed model demonstrated here aims to overcome the problem of imbalanced data, and outperforms common deep learning approaches.

Generalising this work to KAIMRC extended data subset can be limited. Although the extended data subset is newer and richer, discharge type information is not available in this subset.

Epilogue

In this chapter, we investigated the use of Autoencoder models to predict in-hospital mortality risk for patients using the EHR clinical data. A unique approach for handling the data imbalanced was used. The results of the Autoencoder model were compared with commonly used models and approaches for data imbalance handling. In the following chapter, a discussion about the strengths and limitations of this work. Future work will also be presented.

Chapter 9

Discussion

Prologue

This chapter discusses the overall contribution of the previous chapters presented in this thesis and how the predictive models can be applied in healthcare. This includes the limitations related to the medical predictive models employed, the KAIMRC EHR dataset, the population of the dataset, the sampling approaches used, and an evaluation, as well as a discussion of the challenges for generalisation of the findings. Finally, we provide recommendations and directions for further work that is relevant to our thesis topic.

9.1 Discussion

The research questions in this thesis sought to exploit EHR data to facilitate the early identification of patients at risk of the two major medical problems, T2DM and in-hospital mortality using machine learning approaches. Firstly, this involved exploiting EHR data to early identify patients at risk of diabetes via the prediction of Glycated Haemoglobin (HbA1c) levels. This included using EHR data using machine learning in; (i) predicting the levels of Glycated Haemoglobin

(HbA1c) for patients (diabetic and non-diabetic), (ii) identifying patients with pre-diabetes via the prediction of elevated Glycated Haemoglobin (HbA1c) in patients with no history of hyperglycemia, (iii) the use of temporal (time-series) data available in EHR in predicting elevated Glycated Haemoglobin (HbA1c) levels, and (iv) the impact of including temporal behaviour metrics on the importance of variables used to predict elevated Glycated Haemoglobin (HbA1c) levels. Secondly, investigate predicting risk of in-hospital patient mortality machine learning from EHR data.

The results also emphasised that the HbA1c predictive models can exhibit improved learnability when they are employed with the patient longitudinal observations that are normally available from EHR systems. The models employed tended to show better performance when making use of the longitudinal (time-series) behaviour of the predictors. Not only that, utilising patients' longitudinal EHR data affected the relative importance of the predictors used. The models were able to find subtle relationships in the longitudinal data that are relevant to HbA1c prediction, rendering the predictors to be less/more important. Furthermore, including the routinely available variables in the KAIMRC EHR dataset has significantly improved the performance of the models employed. Explainable machine learning methods were used to interrogate the models and provide an understanding of the reasons behind the models' decisions. With regards to in-hospital mortality risk prediction, our findings showed that machine learning achieved promising results using patients clinical data that is routinely stored in the EHR system.

One of the most important success factors for predictive machine learning models is the quality of the data used in training and testing the models (M. Hall and Smith 1998). In this work, we used a large and unique dataset: the KAIMRC EHR dataset, which includes more than 122K unique patients who made records from more than 765K visits to physicians. The KAIMRC dataset contains rich information and a full history of all the visits made by the patients. Most importantly, this dataset includes a clinical diagnosis for patients in terms of diabetes and the discharge type. Also, all clinical observations including laboratory results and interventions were recorded with timestamp details.

Reliable predictive models could be used by health services providers to proactively (as an alarming tool) predict risk score for health conditions/events, such as diabetes or in-hospital mortality, of current and past patients. This can help hospitals, physicians, and patients to plan for in-advance actions. Specifically, it can help hospitals to refine strategic plans for the services to be provided in future. Besides, it can enable clinicians to make quicker patient intervention decisions. Consequentially, the early interventions can benefit patients by minimising delays in diagnosing serious health conditions/complications. The use of the predictive models can help identify the right patient with the right condition at the right time.

One of the main considerations for the successful implementation of predictive model frameworks (machine learning) in healthcare is the continuous adaptation with behavioural changes of predictors and populations used in building the models (Vogenberg 2009). The EHR systems receive large amounts of new data in a regular basis. This data newly-recorded into the EHR systems might contain new patterns/trends or features that are related to the outcomes of the predictive models. Over short/long periods, the predictive models need to be continuously adapted/evaluated to this newly recorded data. With the current advance information technology infrastructures, the process of extracting and moving this data from EHR systems to the predictive model environment can be automated with minimal effort. By automating this process, the EHR systems can continuously supply the predictive models with accurate and up-to-date (real-time) data in a quick and timely manner. Therefore, a dynamic and integrated framework with the EHR systems is important for embarking the use of predictive models that depend on data from EHR systems.

In clinical practice, implementing predictive models is not a one-off task. Proper evaluation the of outcomes of the predictive models is another key component of successful implementation of these models. In particular, this process includes evaluating the inherent benefits/drawbacks of using predictive models in real clinical settings (Harris 2017). Evaluating the performance of the predictive models could be achieved by tracking the interventions that are taken upon the predictive models decisions and comparing those decisions with the actual outcomes (Greenes

et al. 2018). The feedback about those interventions (i.e. followed, match/mismatch) can help evaluating, and consequentially, maintaining/updating the predictive models.

The findings from this work demonstrated that machine learning models can assist in identifying those patients at risk of diabetes/pre-diabetes (via the prediction of HbA1c levels) and at risk of in-hospital mortality from the data available in EHR systems. The machine learning models employed provided reliable performance in these two medical challenges. These challenges were addressed in this thesis by exploiting the KAIMRC dataset using advanced machine learning models for predicting the risk of diabetes/pre-diabetes and in-hospital mortality. Several machine learning technologies (including deep (e.g. MLP, autoencoders, LSTM, BiLSTM and GRU), conventional (e.g. SVM, RF, and LR) and statistical approaches (e.g. MLR)) were employed/investigated to achieve these prediction tasks. In addition, explainable approaches were utilised to expose the reasons behind the classification decisions made by the machine learning forms employed. This was first confirmed by the differentiated replication study described in this thesis. The results obtained from the differentiated replication study also confirmed that different regions in the world can have different weightings of predictors for HbA1c when using predictive models (Wells, Lenoir et al. (2018)'s models specifically) for patients with no history of hyperglycemia. This work also studied effect of incorporating the EHR longitudinal data of the features used into the predictive models employed for prediction of HbA1c elevation levels.

The outcomes of this research project are highly significant: using advanced machine learning models with EHR data would help in enabling clinicians to make quicker patient intervention decisions, to provide improved health-care services to patients at risk of developing diabetes, especially in populations with high diabetes rate such as Saudi Arabia, or at high risk of in-hospital mortality.

With regards to the implication of this work, we hope that our approach can contribute to reducing the massive healthcare costs of diabetes treatment. We also hope that the result of this work can encourage generalising the approaches used in this thesis (advanced machine learning models with EHR data) to address other medical challenges (such as predicting other health conditions or events).

Finally, machine learning models have shown powerful capabilities for analysing and understanding complex data across a wide variety of applications. However, there is still much to do in terms adopting the benefits of using machine learning in the medical domain comparing to other domains such as finance and energy (Harerimana et al. 2019; Panesar 2019). This thesis included pioneering studies that investigated the use of state-of-the-art approaches for diabetes and in-hospital mortality risk predictions. As far as we are aware, this work is the first to study the effect of using routinely collected longitudinal EHR data in the classification performance and use explainability methods to explain the results obtained from the models (for predicting patients with pre-diabetes specifically).

9.2 Limitations

Although EHR systems were designed to improve healthcare outcomes, they were not necessarily intended for research purposes, as mentioned in Chapter 5. Patient data stored in EHR systems can be irregular, as lab instructions are carried out with different frequencies based on the physician's decisions and patient's visit patterns. Commonly, medical data extracted from EHR systems suffer from problems such as irregularity, incompleteness, and noisy and imbalanced data (detailed in Chapter 5). These can be challenging obstacles for any technology used for predictive analytics. Therefore, overcoming such obstacles requires substantial time and effort for data extraction, understanding, and preparation.

Though the use of EHR data in healthcare analytics requires relatively little clinical background, processing and selecting the variables is still challenging. For example, although the machine learning models (and especially the deep learning models) are capable of handling high dimensional data (a large number of variables), using such data does not come without challenges. The more variables that are used, the less complete we expect the records to be. Based on our sampling approach (used in Chapter 7), including variables that have a large number of missing values can make preprocessing highly complicated/time consuming or reduce the number of usable samples overall for training the machine learning models (as observed from the sampling approaches used in this work). Therefore, the theoretical association between the variables

and the outcomes is not the only reason for the inclusion or exclusion of the variables in the employed medical predictive models. To achieve reliable analytics, our analyses imply that it is very important to balance the number of variables involved in the predictive models as well as these variables' availability in the dataset.

EHR systems may contain unwanted data for the predictive models (such as the indicator from clinicians practice). This is because the frequency of laboratory orders made by physicians may add unwanted information to the model. The absence or presence of such predictors can also represent unwanted details in the medical predictive models. To prevent the inclusion of unwanted data, we avoided including data that contains human behavioural indicators. The sampling approach used aimed at minimising including such information as input for the predictive models employed in this thesis by taking the first day data, approximate values in case of multiple patients visits, exclude records with missing values.

Evaluating the predictive models using other datasets can be difficult because of the accessibility and privacy restrictions that apply to other medical datasets. For this reason, and the aforementioned lack of similar studies using machine learning, comparing the performance achieved by the models outlined in this work with those developed by other researchers would require the availability of more suitably anonymised datasets. To overcome this problem, we have split the data using the K fold cross-validation technique to evaluate the predictive models employed. We employed 10 fold cross-validation technique to split the experimental data into training, validation, and testing. The training subset was used to fit the models. The validation subset was used to assess the degree to which a particular model fits into the training subset. In general, the validation subset is used to tune the hyperparameters for the models while the testing subset is used for an unbiased assessment of the final tuned model. We also considered implementing commonly used models (Kopitar et al. 2020), as comparative models, to investigate and compare the results with the employed models.

As we previously highlighted in Chapter 5, data standardisation is a crucial problem in the medical domain. Data standardisation can be an obstacle for validating medical predictive models. Different hospitals use different policies and procedures related to medication, laboratories,

interventions, and so on. As a simple example, the units used for measuring clinical readings can vary from region to region or hospital to hospital, as observed in the replication study in Chapter 7. The different classifications for diagnosis, symptoms, clinical signs and laboratory orders is another example. These differences in the policies and standards followed by hospitals can affect the validation/generalisation of the medical predictive models.

For the KAIMRC dataset specifically, the datasets were collected from the National Guards hospitals in Saudi Arabia. However, these hospitals do not serve the general public. Only those eligible can seek treatment at these hospitals. Thus, the population used in this research is less heterogeneous than the general population with regard to the inclusion of different ethnic groups. Besides, the data used for training and testing the predictive models was collected from Saudi Arabia; a population that has a high risk of T2DM, and so the findings may not readily generalise to other countries. The diabetes rate in the KAIMRC dataset (collected from Saudi Arabia) is considerably higher compared to other datasets such as that used by Wells, Lenoir et al. (2018).

A further limitation is the absence of certain variables in some EHR datasets. There are many reasons for the absence of variables; for example, clinician's decisions and the cost of particular tests. In addition, geographical factors can affect the inclusion of particular variables. For example, in some regions, it is very important for most patients visiting hospitals to be routinely checked for several common heath conditions that have high prevalence rates in these regions. As an example, the possibility of a patient being checked against hyperglycemia is high in the population of Saudi Arabia, due to diabetes prevalence (Al-Zahrani et al. 2019). Therefore, we aimed at investigating mining the available data in the KAIMRC EHR system.

Since the KAIMRC dataset contains structured numerical/categorical (time-series) data, we focused on applying machine learning approaches that are applicable to, and commonly used with, the structured data (RF, MLR, LR, SVM, MLP, LSTM, BiLSTM, GRU, and autoencoder).

Finally, in clinical trials, unobserved clinical values are never known precisely (Cro et al. 2020). Therefore, filling in missing values for medical data requires proper clinical justifications. Most

importantly, missing data imputation in medical datasets can not only impact models' classification performance but also the importance of relative predictors in predictive models (Payrovnaziri et al. 2020). For these reasons, we avoided using intensive features engineering in all sampling approaches used in this research. As seen in the sampling approaches used in Chapter 7, rather than applying techniques for missing data imputations, we chose to eliminate records that have one or more missing values and operate on only the complete data records (Jazayeri, Liang and C. C. Yang 2020).

9.3 Future Work

Based on our findings in this work, it could be valuable to explore the impact of applying sampling approaches in future work, as would exploration of the use of a larger dataset with more variables. Unfortunately such variables as genetic makeup or lifestyle characteristics (Elhadd, Al-Amoudi and Alzahrani 2007) are difficult to collect and incorporate into the EHR systems. However, the availability of such variables may help to better account for the outcomes of this research project.

If this study were to be repeated, we would prioritise the gathering of extra data to improve our predictive models. More specifically, we would prioritise details such as the patient's family history of diabetes, their smoking status (including for how long they have been a smoker), their job demands (considering physical activity and working hours), current and/or previous health conditions, medications, and the patient's history of pregnancy (for women). We postulate that incorporating these details (in addition to laboratory tests and vital signs) would help to improve the identification of patients at risk of diabetes or pre-diabetes using predictive models.

We would also investigate applying other machine learning approaches (conventional and deep) such as ELM (discussed in the related work chapter) and Hidden Markov Models (Perveen et al. 2019). Another direction would be generalising the Col-DAE used in Chapter 6 for other prediction tasks using larger datasets. Examining combining multiple classifiers in one predictive model (framework) would be another option for future work.

Furthermore, we would also seek to incorporate longitudinal behaviour of the predictors (laboratory tests and vital signs) over longer periods into the models used in this thesis. We believe that including more variations of the values for the predictor's between a patient's visits would increase the model's prediction efficiency.

Regardless of the drawbacks of the imputation approaches for the missing data, studying the impact of applying other methods for handling the missing data is another component of the future work related to this research project's results. This may include studying the impact of the different missing data interpretation approaches on the feature's importance using the employed models. Another direction for the future work would be investigating applying machine learning for predicting other medical health conditions using EHR data and/or investigating the performance of the models employed by using different populations.

Epilogue

This chapter has provided a discussion about the results obtained from Chapters 6, 7 and 8. We highlighted the main outcomes of the studies presented in these chapters. We identified the limitations and challenges of adopting state-of-the-art machine learning approaches using EHR data. The following chapter will draw an overall conclusion of the this thesis.

Chapter 10

Conclusion

This thesis provides several key contributions to the area of healthcare informatics. We have explored how state-of-the-art machine learning approaches can be used to address two key challenges in healthcare using EHR data: (i) identifying patients at risk of developing diabetes and (ii) identifying in-hospital mortality. Given the difficulties inherent in employing EHR data, this thesis has shown that our machine learning models employed were able to achieve promising results for these two challenges.

We considered the large and unique KAIMRC dataset that contains a full history of patients' visits. We began by addressing the challenge of predicting a clinical diagnosis of diabetes using data that are routinely collected and stored in hospital EHR systems. We investigated the diagnoses of patients with diabetes via the prediction of HbA1c levels in patients' blood. To achieve this, we used a novel collaborative denoising autoencoder (Col-DAE) framework. This framework built an independent denoising autoencoder model for high and low HbA1c levels, which extracts feature representations in the latent space. The latent spaces of both models are then merged and passed to a MLP model for decision-making. The framework demonstrated

that a patient's risk of developing high levels of HbA1c can reliably be predicted from EHR records.

Then, we have addressed the challenge of identifying patients with pre-diabetes. First, we investigated performing a differentiated replication study to validate, evaluate, and identify the strengths and weaknesses of replicating a predictive model on different population. This model employed multiple logistic regression with EHR data to forecast elevated levels of HbA1c for patients with no history of diabetes. The study being replicated used data from a population from the United States and the differentiated replication used data from a population from Saudi Arabia. The study showed that the direct use of the models (calculators) created using multiple logistic regression to predict the level of HbA1c may not be appropriate for all populations. This particular study also revealed that the weighting of the predictors needs to be calibrated to the population used. However, the study did confirm that replicating the original study using a different population can help with predicting the levels of HbA1c by using the predictors that are routinely collected and stored in hospital EHR systems.

As an extension of the replication study, this thesis investigated the performance of predictive models to forecast HbA1c elevation levels, to identify patients with pre-diabetes. This was achieved by employing machine learning approaches using data from current and previous visits (longitudinal data) stored in EHR systems for patients who had not been previously diagnosed with any type of diabetes. To add a degree of transparency, explainable methods been used to interpret the decision made by the blackbox models employed. Explainable methods were used to rank the relative importance of the features used. Following that, we carried out more investigation into the predictability of current elevated HbA1c levels using more of the data provided in EHR datasets that are normally available.

Our finding showed that machine learning models can provide promising results for the task of predicting current HbA1c levels. The models, including the statistical ones, utilising the patient's longitudinal time-series data improved the performance and affected the relative importance of the predictors used. The results also showed a significant improvement in the performance

of the employed models when including more features routinely available in the EHR systems highlighting the importance of extra data/large datasets. For the MLP model, the accuracy jumped from 73.57% to 74.51% using longitudinal EHR data, and up to 77.37% by incorporating larger dimensional EHR data. This could be significant for clinicians when seeking to improve the healthcare services provided for patients and reduce the associated cost.

Up to our knowledge, this work has been the first to employ deep learning approaches in predicting the HbA1c elevation levels using the data extracted from the EHR systems. It was also the first to investigate the effect of the EHR longitudinal data and to use explainable methods to explain the decisions of the models used for the HbA1c levels prediction task.

We also investigated a novel application of the Stacked Denoising Autoencoder (SDA) for in-hospital patients mortality risk prediction. We uniquely tackled the problem of data imbalance that usually exists in this area of research. The model was trained using only the majority class data of the KAIMRC dataset. It was tested using a mixture of majority and minority classes. The employed model, using patient's clinical data from a variety of health conditions and without intensive feature engineering, achieved robust and promising results using only the first 24 hours of patient data after admission. This confirms that the EHR data with the advanced machine learning models can assist in predicting patients at risk of in-hospital mortality. Accurate assessment of in-hospital mortality is highly significant for clinicians to allow for preventive interventions that can help improve a patient's likelihood of survival.

To conclude, we believe that the outcomes of this research project are highly significant, enabling clinicians to make quicker patient intervention decisions, to provide improved health-care services to patients at risk of developing diabetes or at high risk of in-hospital mortality. The early identification of diabetes can help benefit patients and clinicians to plan for preventive interventions that can delay and/or prevent serious complications. We hope that our approach can contribute to improving patient's survival and reducing the massive healthcare costs of diabetes treatment.

Appendix A

Supplementary about the Structure of KAIMRC Dataset

A.1 KAIMRC Data Part 1

Below is structure details for the patient, clinical diagnosis, vital signs, and lab test files collected for KAIMRC dataset part 1 (all collected in one file structure):

Table A.1: The structure of the data collected from KAIMRC part 1.

Field	**Description**	**Type**
MRN	Patient Medical Record Number	Integer number
VISIT_ID	Patient visit ID	Integer number
VISIT_NUM	Visit number	Text
VISIT_TYPE	Visit type	Text (Categorical)
VISIT_STATUS	Visit status	Text (Categorical)
NATIONALITY	Nationality of patient	Text (Categorical)
GENDER	Gender of patient	Text (Categorical)

Table A.2: The structure of the data collected from KAIMRC part 1 (cont).

Field	**Description**	**Type**
CURRENT_AGE	Current age of patient	Integer number
VISIT_AGE	Patient age at visit time	Integer number
WEIGHT	Weight of patient	Integer number
HEIGHT	Height of patient	Integer number
BMI	Body Mass Index	Integer number
HBP	High Blood Pressure (systolic)	Integer number
LBP	Low Blood Pressure (diastolic)	Integer number
REGION_NAME	Region	Text (Categorical)
HGB_A1C	Glycated haemoglobin	Float number
DIAGNOSIS	Diabetes clinical diagnosis	Text (Categorical)
FACILITY	Hospital name	Text (Categorical)
ADMISSION_DATE	Admission date	Date
ADMISSION_DAY	Day of admission	Integer number
ADMISSION_MONTH	Month of admission	Integer number
ADMISSION_YEAR	Year of admission	Integer number
ADMISSION_TIME	Admission time	Time
ADMISSION_AM_PM	Admission meridiem period	Text (Categorical)
DISCHARGE_DATE	Date of patient discharge	Date
DISCHARGE_DAY	Day of patient discharge	Integer number
DISCHARGE_MONTH	Month of patient discharge	Integer number
DISCHARGE_YEAR	Year of patient discharge	Integer number
DISCHARGE_TIME	Discharge time	Time
DISCHARGE_AM_PM	Discharge meridiem period	Text (Categorical)
DISCHARGE_TYPE	Discharge type	Text (Categorical)
LOS	Length of Stay	Integer number
SERVICE_NAME	Service name provided	Text (Categorical)
SOURCE	Source of service	Text (Categorical)
DEPARTMENT	Department of service	Text (Categorical)
SPACILTY	Spacilty	Text (Categorical)
UNIT	Unit of the service	Text (Categorical)
BEDTYPE	Bed type	Text
PROCEDURE	Clinical procedure for the test	Text
PROFILE	Test profile	Text (Categorical)
LAB_DATE	Laboratory test date	Date
TEST	Laboratory test name	Text (Categorical)
VALUE	Laboratory test value	Text
ABNORMAL_SATE	Abnormality status	Text (Categorical)
ALLERGY_ASSESSMENT	Assessment of patient allergies	Text
DIAGNOSIS	Clinical diagnosis of diabetes	Text (Categorical)

A.2 KAIMRC Data Part 2

Below are structure details for the patient , clinical diagnosis, vital signs and lab test files collected for dataset part 2 (collected in several file with different structures):

Table A.3: The structure of the data collected from KAIMRC part 2 for patient file.

Field	**Description**	**Type**
REGION	Region	Text (Categorical)
FACILITY	Hospital name	Text (Categorical)
ETPR_PT_NO	Patient Electronic Therapeutic Plan Register number	Integer number
PT_NO	Patient medical number	Integer number
GENDER	Gender of patient	Text (Categorical)
BIRTHDAY_DATE	Patient birthday date	Date
DECEASED	Patient diabetes clinical diagnosis	Character (Categorical)
DEATH_DATE	Patient death date	Date

Table A.4: The structure of the data collected from KAIMRC part 2 for patient clinical diagnosis file.

Field	**Description**	**Type**
ETPR_PT_NO	Patient Electronic Therapeutic Plan Register number	Integer number
MRN	Patient Medical Record Number	Integer number
VISIT_NUM	Visit number	Text
DIAGNOSIS	Clinical diagnosis description	Text
ICD10_CD	Diagnosis ICD10 code	Text (Categorical)
DIAGNOSIS_DATE	Diagnosis date	Date
MAIN_SICKNESS	Main sickness	Character (Categorical)

Table A.5: The structure of the data collected from KAIMRC part 2 for patient laboratory test file.

Field	**Description**	**Type**
REGION	Region	Text (Categorical)
FACILITY	Hospital name	Text (Categorical)
ETPR_PT_NO	Patient Electronic Therapeutic Plan Register number	Integer number
MRN	Patient Medical Record Number	Integer number
PACT_ID	PACT Visit number	Text
EXM_CD	Laboratory test ICD10 code	Text (Categorical)
EXM_NM	Laboratory test description	Text
BRFG_DTM	Laboratory test date	Date
RESULT	Laboratory test result	Float number

Table A.6: The structure of the data collected from KAIMRC part 2 for patient vital signs files.

Field	**Description**	**Type**
ETPR_PT_NO	Patient Electronic Therapeutic Plan Register number	Integer number
MRN	Patient Medical Record Number	Integer number
VISIT_NUM	Visit number	Text
HEIGHT	Height of patient	Integer Number
WEIGHT	Weight of patient	Integer Number
DATE_	Vital sign date	Date

Appendix B

Laboratories Tests in KAIMRC Dataset

Table B.1: Laboratory tests used in KAIMRC (parts 1 and 2).

Laboratory Test Description	**ICD10 Code**	**Used in Subset**
Urea Nitrogen, Blood (BUN)	L3900233	A, D, E
Creatinine Level, Serum (Crea)	L3900104	A, D, E
Sodium Level, Serum (Na Lvl)	L2000011	A, D, E
Haematocrit (Hct)	L3900230	A, D, E
Haemoglobin (Hgb)	L2000009	A, D, E
Chloride Level, Serum (Cl Level)	L3900234	D
Mean Cell Haemoglobin Concentration (MCHC)	L2000010	A, D, E
Mean Cell Volum (MPV)	L2000018	A, D, E
White Blood Cell Count (WBC Count)	L2000016	A, D, E
Carbon Dioxide Level (CO2)	L2000014	A, D, E
Red Blood Cells width (RDW)	L3900231	A, D, E
Platelet Count (Plt Count)	L2000006	A, D, E
Mean Corpuscular Volume (MCV)	L2000015	A, D, E
Red Blood Cell (RBC)	L2000008	A, D, E
Mean Cell Haemoglobin (MCH)	L2000012	A, D, E

Table B.2: Laboratory tests used in KAIMRC (parts 1 and 2) (cont.).

Laboratory Test Description	**ICD10 Code**	**Used in Subset**
estimated Glomerular Filtration Rate (eGFR)	L3000002	A, B, D, E
Anion Agap (AGAP)	L3900102	A, D, E
Total Bilirubin (T Bili)	L3000012	A, D, E
Glucose Level, Random (RBS, Glur)	L3900229	A, B, C, D, E
Cholesterol Level (CHOL)	L3000006	A, B, C, D, E
High Density Lipoprotein (HDL)	L3000013	A, B, C, D, E
Triglyceride Level (Trig)	L3000019	A, D, E
Low Density Lipoprotein (LDL)	L3000026	A, B, C, D, E
Potassium Level, Serum (K Level, Pota)	L3900232	A, E
Aspartate Aminotransferase (SGOT)	L3900111	A, E
Nucleated Red Blood Cells (NRBC%)	L2000013	A, E
Alanine Transferase (SGPT)	L3900109	A, E
Total Protein, Serum (TP)	L3000022	A, E
Alkaline Phosphatase Level (Alk Phos)	L3000018	A, E
Albumin Level, Serum (Albu)	L3000014	A, E
Adjusted Calcium (AdCa)	L3000027	A, E
Calcium (Ca)	L3900208	A, E
Phosphorus Level (P04 Lvl, Phos)	L3000008	A, E
Magnesium Level, Serum (Mg)	L3000007	A, E
Uric Acid, Serum (UrAc)	L3000011	A, E
Glucose Level, Fasting (FBS)	L3900113	D
Glycated Haemoglobin (HbA1c)	L3700025	E
Thyroid Stimulating Hormone (TSH)	L3200010	A, E
Protime Time	L7100108	A, E
International Normalized Ration (INR)	L7100109	A, E
Partial Prothrombin Time (PTT)	L7100110	A, E
UA Spec Grav	L6100024	A, E
Troponin I Level	L3500019	A, E
Creatinin kinase (CK, CPK)	L3000023	A, E
B-Type Natriuretic Peptide (BNP)	L3900056	A, E
UA WBC	L6100030	A, E
UA Protein	L6100026	A, E

Table B.3: Laboratory tests used in KAIMRC (parts 1 and 2) (cont.).

Laboratory Test Description	**ICD10 Code**	**Used in Subset**
UA RBC	L6100029	A, E
Free Thyroxine Level (Free T4)	L3900178	A, E
Gamma Glutamyl Transferase (GTP)	L3200006	A, E
UA Blood	L6100025	A, E
UA Leuk Est	L6100028	A, E
C-Reactive Protein (CRP)	L5100170	A, E
UA Glucose	L6100022	A, E
UA Ketones	L6100023	A, E
Immunoglobulin G	L5100001	A, E
Others Manual Blood Differential Test#	L2000052	A, E
Others Manual Blood Differential Test%	L2000039	A, E
Thrombin Time	L7100062	A, E
Nucleated Red Blood Cells (NRBC#)	L2000017	A, E
Random Urine Creatinine	L3100045	A, E
Vancomycin Level, Trough	L3900129	A, E
UA Squam Epithelial	L6100012	A, E
Ldlc Coronary Risk (JD)	NA	A, E
Alb from Renal Profile	NA	A, E
Mg from Renal Profile	NA	A, E
Blood Type Test (Abroh)	NA	A, E
Antibody Screen (D Event)	NA	A, E
Neutropenia Auto #	L2700005	A, E
Monocytes Auto #	L2700004	A, E
Basophils Auto #	L2700007	A, E
Eosinophil Count #	L2700003	A, E
Lymph Auto #	L2700012	A, E
Monocytes Auto %	L2700008	E
Eosinophil Auto %	L2700009	E
Neutropenia Auto %	L2700010	E
Basophils Auto %	L2700014	E
Lymph Auto %	L2700013	E
Direct Bilirubin (Conjugated Bilirubin)	L3000010	E
Ca from Renal Profile	NA	E

- The abbreviations between the parenthesis, in some cases are used for shortening the name of the laboratory test for presentation purposes (not necessarily the abbreviations used in the medical practice).
- Few laboratory tests in dataset part 1 were not available in dataset part 2 hence the missingness ICD code.
- Unused laboratory tests in this thesis were not listed.

Appendix C

Hyperglycemia ICD10 Diagnostic Codes

Table C.1: ICD10 Hyperglycemia diagnostic codes used by KAIMRC.

Diagnostic ICD10 Code	**Description**
E11	Type 2 Diabetes Mellitus (T2DM)
E14	Diabetes Mellitus
E10	Type 1 Diabetes Mellitus
E139	Familial diabetes mellitus
R73	Hyperglycemia
O24	Gestational diabetes

Appendix D

Units Conversion

Laboratories Units Conversion Details

Total Cholesterol (CHOL) and Non-High Density Lipoprotien (non-HDL)[1]

To convert from mmol/L to mg/dL: Multiply by 38.67

To convert from mg/dL to mmol/L: Multiply by 0.02586.

Random Blood Sugar (Glucose) (RBS) and Fasting Blood Sugar (Glucose) (FBS)[2]

To convert from mmol/L to gm/dL: Multiply by 18

To convert from gm/dL to mmol/L: Multiply by 0.01129

[1] Reference: https://www.ncbi.nlm.nih.gov/books/NBK33478/
[2] Reference: https://www.diabetes.co.uk/blood-sugar-converter.html#

Appendix E

PM Calculator

Table E.1: PM3 Calculator details for predicting HbA1c level.

Intercept		
	–	4.5404143
Random Blood Sugar (Glucose) Level (RBS)		
	–	0.010264092 * RBS
	+	5.9135831e-05 * max(RBS - 79.2,0)^3
	–	0.00015189117 * max(RBS - 93.6,0)^3
	+	0.00010465775 * max(RBS - 106.2,0)^3
	–	1.1784309e-05 * max(RBS - 131.4,0)^3
	–	1.181092e-07 * max(RBS - 277.2,0)^3
estimated Glomerular Filtration Rate (eGFR)		
	+	0.040416671* eGFR
	–	5.6881504e-06 * max(eGFR - 15,0)^3
	+	4.98899e-05 * max(eGFR - 74,0)^3
	–	8.7199708e-05 * max(eGFR - 92,0)^3
	+	4.8759935e-05 * max(eGFR - 105.92507,0)^3
	–	5.761977e-06 * max(eGFR - 130,0)^3
Age		
	+	0.074906466 * Age
	–	7.6062958e-06 * max(Age - 25,0)^3
	–	1.1470552e-05 * max(Age - 44,0)^3
	–	2.2987067e-05 * max(Age - 55,0)^3
	+	9.3130665e-05 * max(Age - 66,0)^3
	–	5.106675e-05 * max(Age - 82,0)^3
Total Cholesterol (CHOL)		
	–	0.021457758 * CHOL
	–	1.2304102e-06 * max(CHOL - 105.1824,0)^3
	+	5.0953941e-06* max(CHOL - 138.8253,0)^3
	–	5.9498557e-06* max(CHOL - 163.5741,0)^3
	+	2.1424285e-06* max(CHOL - 191.0298,0)^3
	–	5.7556768e-08* max(CHOL - 242.8476,0)^3
Non-High Density Lipoprotien (non-HDL)		
	+	0.034382335 * non-HDL
	–	5.0345293e-06 * max(non-HDL - 72.3129,0)^3
	+	1.4439644e-05 * max(non-HDL - 99.3819,0)^3
	–	1.1919275e-05 * max(non-HDL - 120.2637,0)^3
	+	2.6133174e-06 * max(non-HDL - 146.1726,0)^3
	–	9.915721e-08 * max(non-HDL - 196.8303,0)^3

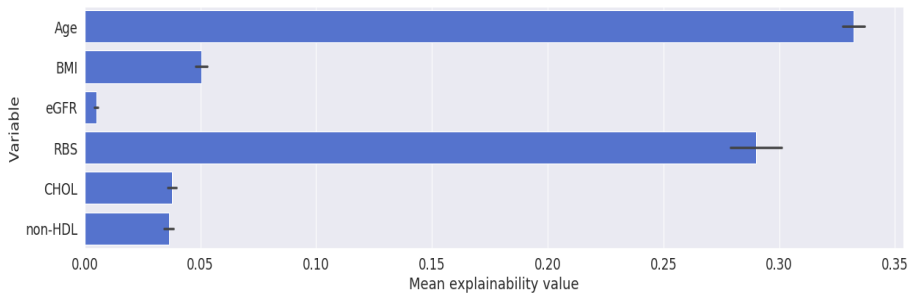
Appendix F

Predictors Relative Importance Charts

Predictors relative importance charts for the models

Multiple Logistic Regression (MLR)

Figure F.1: Relative importance of predictors for the MLR.



Random Forest (RF)

Figure F.2: Relative importance of predictors for the RF model trained without longitudinal data.

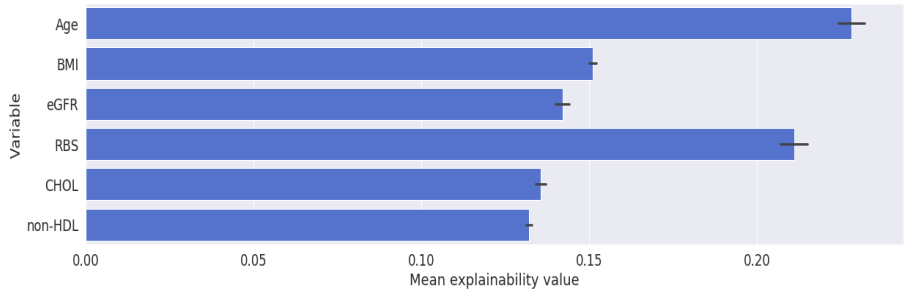
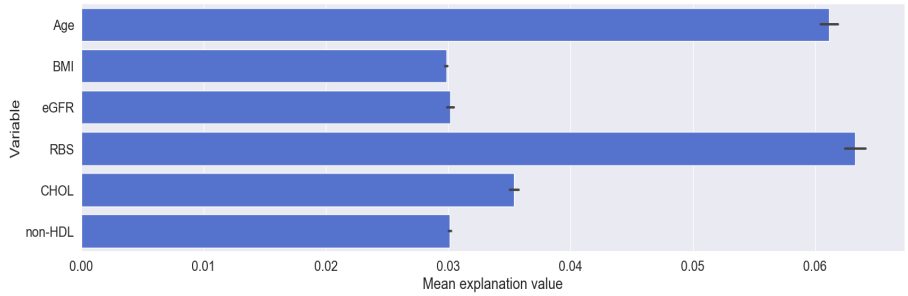


Figure F.3: Relative importance of predictors for the RF model trained with longitudinal data.



Logistic regression (LR)

Figure F.4: Relative importance of predictors for the LR model trained without longitudinal data.

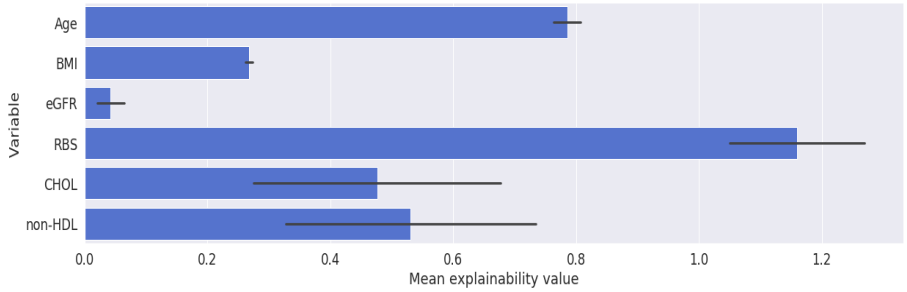
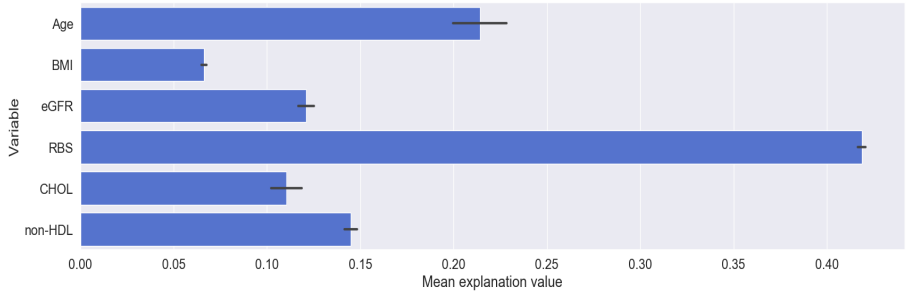


Figure F.5: Relative importance of predictors for the LR model trained with longitudinal data.



Support Vector Machine (SVM)

Figure F.6: Relative importance of predictors for the SVM model trained without longitudinal data using SHAP.

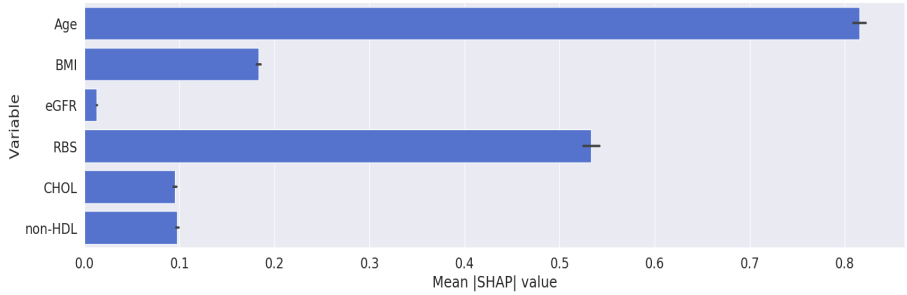


Figure F.7: Relative importance of predictors for the SVM model trained without longitudinal data using LIME.

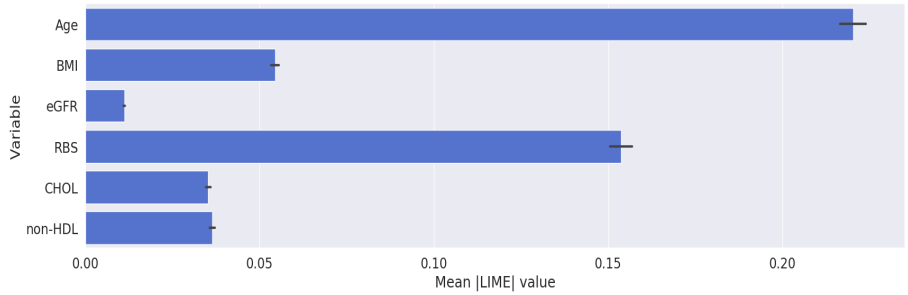


Figure F.8: Relative importance of predictors for the SVM model trained with longitudinal data using SHAP.

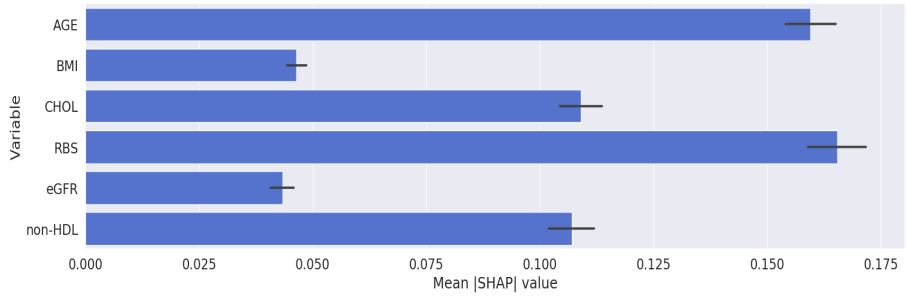
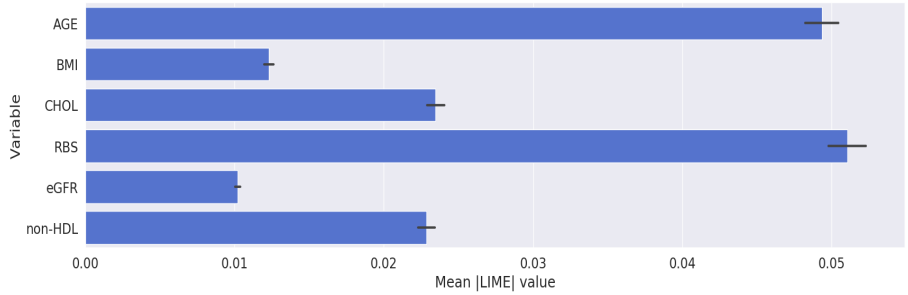


Figure F.9: Relative importance of predictors for the SVM model trained with longitudinal data using LIME.



Multi-layer perceptron (MLP)

Figure F.10: Relative order of importance of predictors for the MLP model trained without longitudinal data using SHAP.

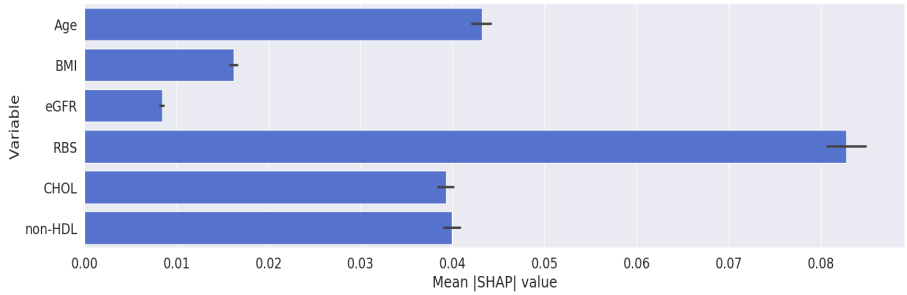


Figure F.11: Relative order of importance of predictors for the MLP model trained without longitudinal data using LIME.

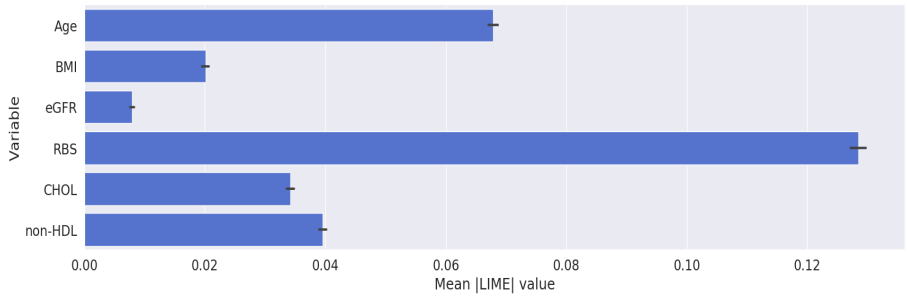


Figure F.12: Relative order of importance of predictors for the MLP model trained with longitudinal data using SHAP.

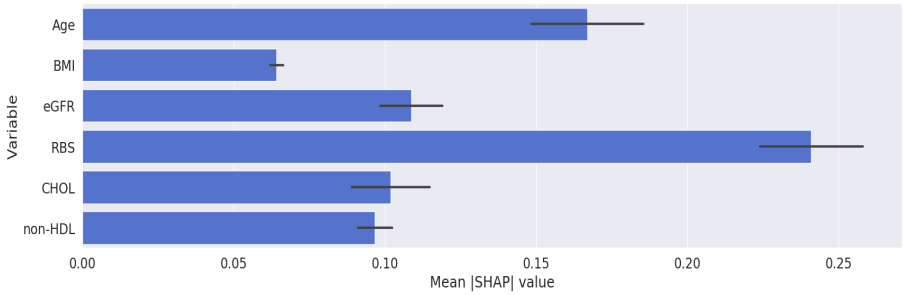
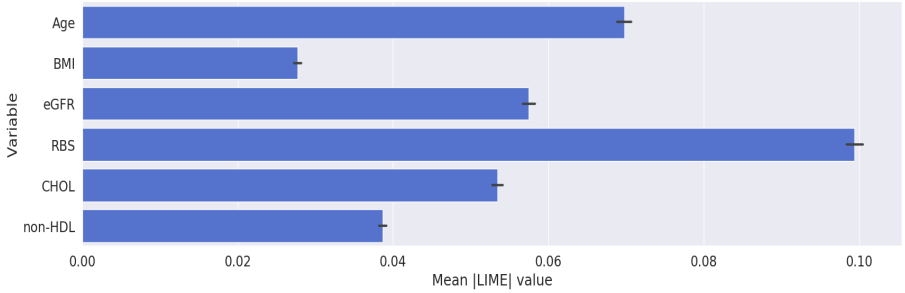


Figure F.13: Relative Order of importance of predictors for the MLP model trained with longitudinal data using LIME.



Appendix G

LR and MLR Calculators

Table G.1: The details of LR trained without longitudinal data used for predicting HbA1c elevation levels using data subset(C).

Intercept		
	+	0.23815679
Random Blood Sugar (Glucose) Level (RBS)		
	+	1.06104915
estimated Glomerular Filtration Rate (eGFR)		
	−	0.0521671
BMI		
	+	0.26236068
Age		
	+	0.75756277
Total Cholesterol (CHOL)		
	−	0.32541754
Non-High Density Lipoprotien (non-HDL)		
	+	0.36907342

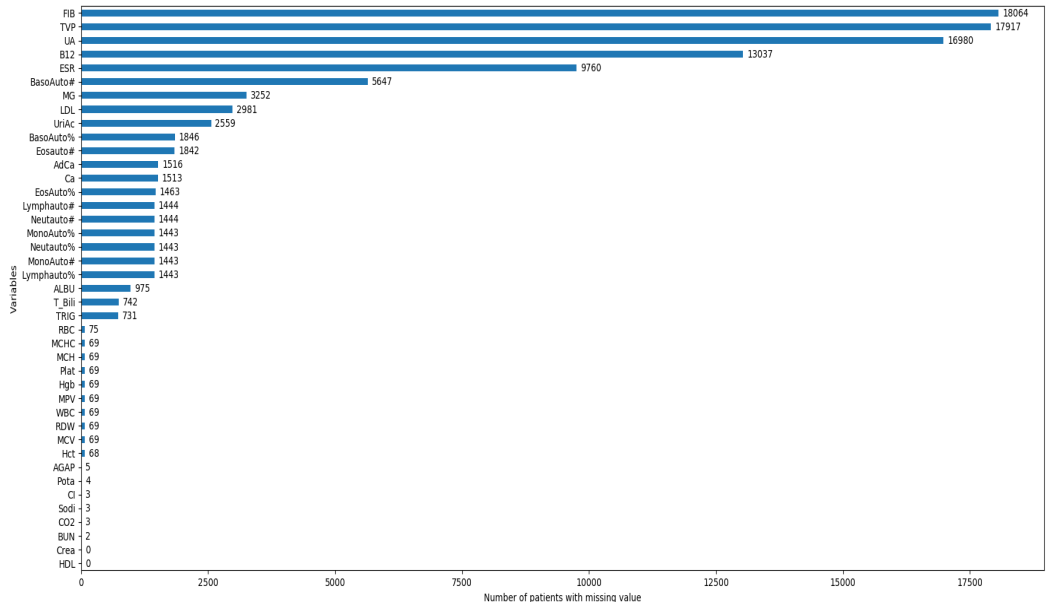
Table G.2: The details of MLR trained without longitudinal data used for predicting HbA1c Elevation levels using data subset(C).

Intercept		
	−	7.4910884
Random Blood Sugar (Glucose) Level (RBS)		
	+	0.47325147 * RBS
	−	0.0080969994 * max(RBS - 4.6,0)^3
	+	0.0099146931 * max(RBS - 5.7,0)^3
	−	0.0018176937 * max(RBS - 10.6,0)^3
estimated Glomerular Filtration Rate (eGFR)		
	+	0.0.0048805867 * eGFR
	−	1.1069511e-06 * max(eGFR - 51,0)^3
	+	2.8246337e-06 * max(eGFR - 96,0)^3
	−	1.7176827e-06 * max(eGFR - 125,0)^3
BMI		
	+	0.055349506 * BMI
	−	8.3374022e-05 * max(BMI - 22.41,0)^3
	+	0.00014121475 * max(BMI - 29.07,0)^3
	−	5.7840728e-05 * max(BMI - 38.67,0)^3
Age		
	+	0.084259781 * Age
	−	2.2417323e-05 * max(Age - 29,0)^3
	+	4.3041261e-05 * max(Age - 52,0)^3
	−	2.0623938e-05 * max(Age - 77,0)^3
Total Cholesterol (CHOL)		
	−	0.77645878 * CHOL
	+	0.0086131708 * max(CHOL - 3.11,0)^3
	−	0.01640604 * max(CHOL - 4.44,0)^3
	+	0.0077928688 * max(CHOL - 5.91,0)^3
Non-High Density Lipoprotien (non-HDL)		
	+	0.71595282 * non-HDL
	+	0.0071327364 * max(non-HDL - 2.15,0)^3
	−	0.01264216 * max(non-HDL - 3.27,0)^3
	+	0.0055094239 * max(non-HDL - 4.72,0)^3

Appendix H

Missingness in KAIMRC

Figure H.1: Number of patients with missing values for the available variables in KAIMRC data subset(D).

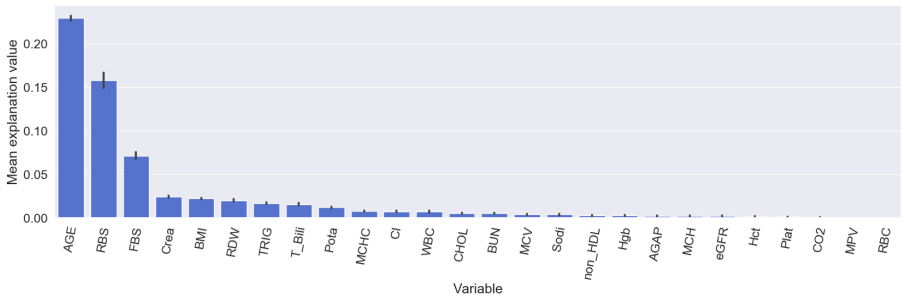


Appendix I

Predictors Relative Importance For Models Using Top Available Variables

Multiple Logistic Regression (MLR)

Figure I.1: Relative importance of predictors for the MLR.



Random Forest (RF)

Figure I.2: Relative importance of predictors for the RF without longitudinal data.

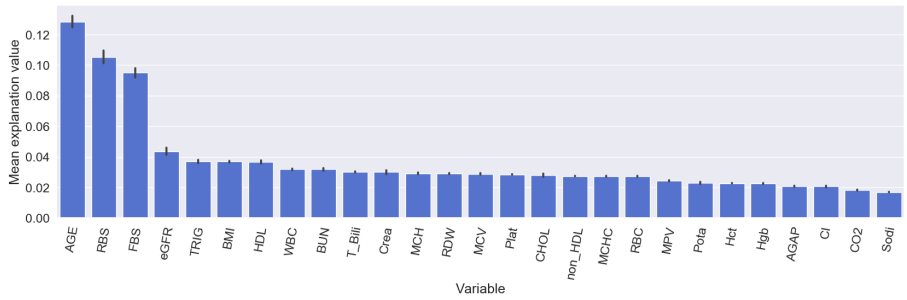


Figure I.3: Relative importance of predictors for the RF without longitudinal data using SHAP.

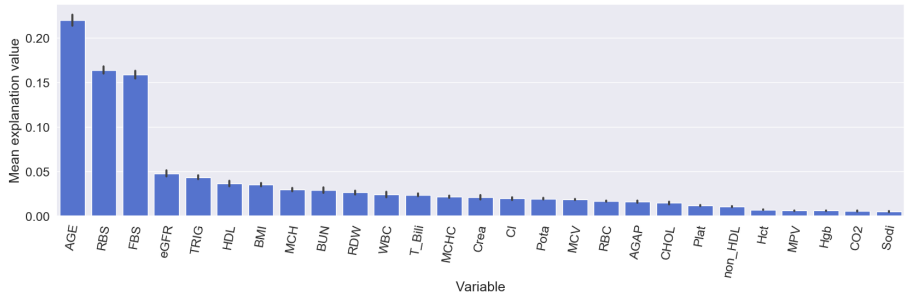


Figure I.4: Relative importance of predictors for the RF with longitudinal data.

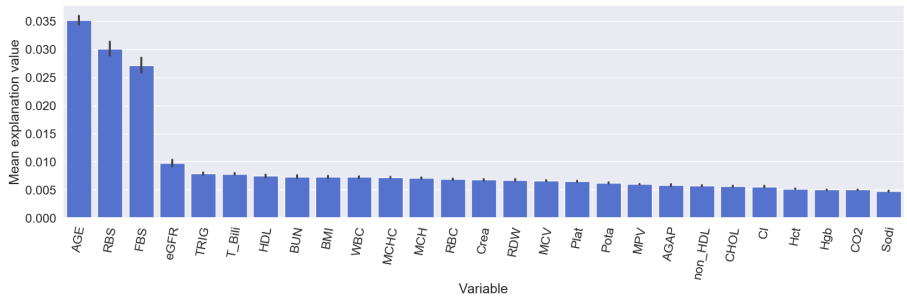
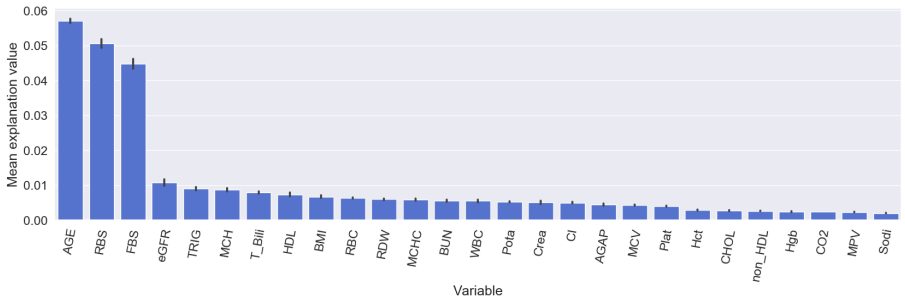


Figure I.5: Relative importance of predictors for the RF with longitudinal data using SHAP.



Support Vector Machine (SVM)

Figure I.6: Relative importance of predictors for the SVM without longitudinal data using SHAP.

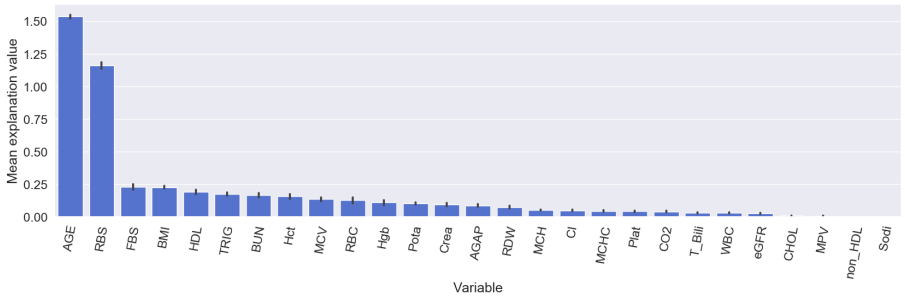
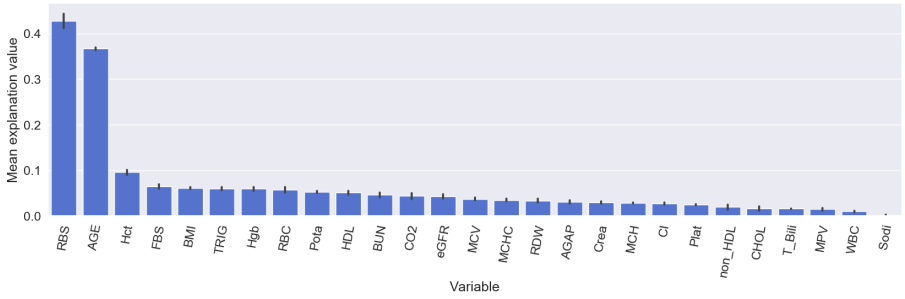


Figure I.7: Relative importance of predictors for the SVM with longitudinal data using SHAP.



Logistic Regression (LR)

Figure I.8: Relative importance of predictors for the LR without longitudinal data.

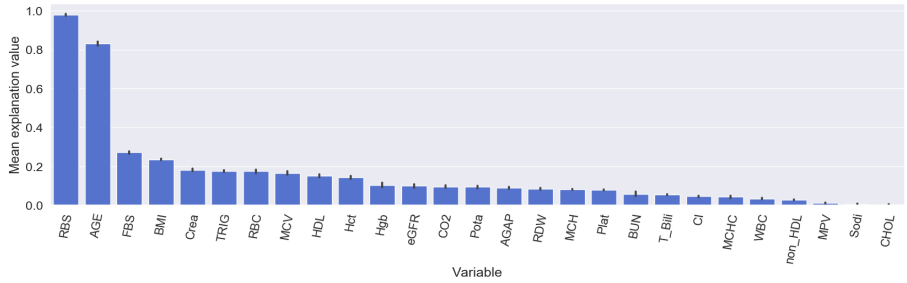
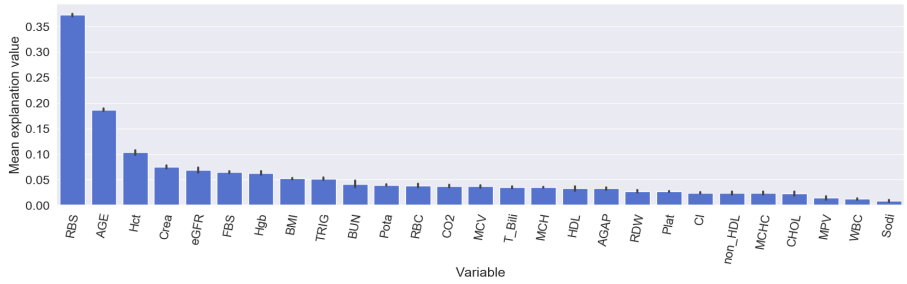


Figure I.9: Relative importance of predictors for the LR with longitudinal data.



Multi-layer perceptron (MLP)

Figure I.10: Relative importance of predictors for the MLP without longitudinal data using SHAP.

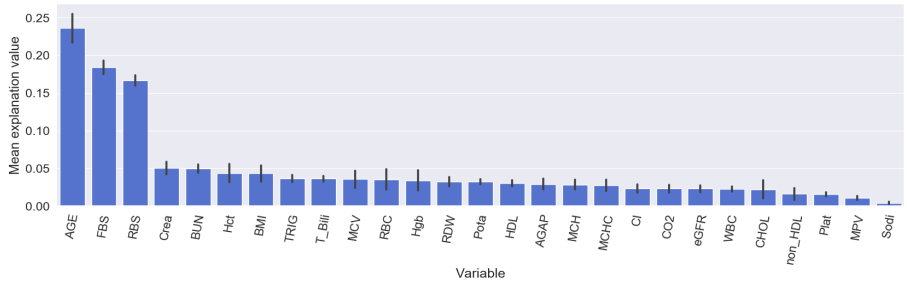
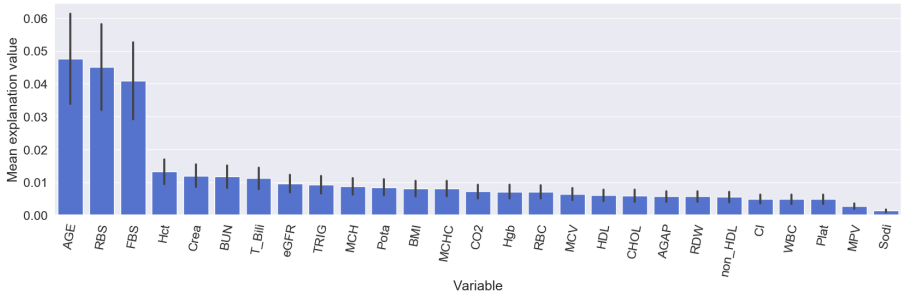


Figure I.11: Relative importance of predictors for the MLP with longitudinal data using SHAP.



Bibliography

Abdulaziz Al Dawish, Mohamed et al. (2016). 'Diabetes mellitus in Saudi Arabia: a review of the recent literature'. In: *Current diabetes reviews* 12.4, pages 359–368 (*pp.* 7, 67, 68, 106).

Abhyankar, Swapna, Dina Demner-Fushman and Clement J McDonald (2012). 'Standardizing clinical laboratory data for secondary use'. In: *Journal of biomedical informatics* 45.4, pages 642–650 (*p.* 64).

Abouelmehdi, Karim, Abderrahim Beni-Hessane and Hayat Khaloufi (2018). 'Big healthcare data: preserving security and privacy'. In: *Journal of Big Data* 5.1, page 1 (*p.* 64).

Ackermann, Ronald T et al. (2011). 'Identifying adults at high risk for diabetes and cardiovascular disease using hemoglobin A1c: National Health and Nutrition Examination Survey 2005–2006'. In: *American journal of preventive medicine* 40.1, pages 11–17 (*pp.* 2, 8, 18, 19, 21, 82).

Ahmad, Muhammad Aurangzeb, Carly Eckert and Ankur Teredesai (2018). 'Interpretable machine learning in healthcare'. In: *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 559–560 (*p.* 111).

Aishwarya, R, P Gayathri et al. (2013). 'A method for classification using machine learning technique for diabetes'. In: (*pp.* 38, 49, 51).

Al Moubayed, Noura et al. (2016). 'Sms spam filtering using probabilistic topic modelling and stacked denoising autoencoder'. In: *International Conference on Artificial Neural Networks*. Springer, pages 423–430 (*p.* 83).

Alessa, Ali, Miad Faezipour and Zakhriya Alhassan (2018). 'Text classification of flu-related tweets using fasttext with sentiment and keyword features'. In: *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, pages 366–367 (*p.* viii).

Alhassan, Zakhriya, David Budgen, Ali Alessa et al. (2019). 'Collaborative Denoising Autoencoder for High Glycated Haemoglobin Prediction'. In: *International Conference on Artificial Neural Networks*. Springer, pages 338–350 (*pp.* vi, 82).

Alhassan, Zakhriya, David Budgen, Riyad Alshammari and Noura Al Moubayed (2020). 'Predicting Current Glycated Hemoglobin Levels in Adults From Electronic Health Records: Validation of Multiple Logistic Regression Algorithm'. In: *JMIR medical informatics* 8.7, e18963 (*pp.* vi, 93).

Alhassan, Zakhriya, David Budgen, Riyad Alshammari, Tahani Daghstani et al. (2018). 'Stacked denoising autoencoders for mortality risk prediction using imbalanced clinical data'. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, pages 541–546 (*pp.* vi, 139).

Alhassan, Zakhriya, A Stephen McGough et al. (2018). 'Type-2 diabetes mellitus diagnosis from time series clinical data using deep learning models'. In: *International Conference on Artificial Neural Networks*. Springer, pages 468–478 (*p.* v).

Alhassan, Zakhriya, Matthew Watson et al. (2021). 'Improving Current Glycated Hemoglobin Prediction in Adults: Use of Machine Learning Algorithms With Electronic Health Records'. In: *JMIR Medical Informatics* 9.5, e25237 (*p.* 93).

Ali, Peshawa Jamal Muhammad et al. (2014). 'Data normalization and standardization: a technical report'. In: *Mach Learn Tech Rep* 1.1, pages 1–6 (*p.* 144).

Alpaydin, Ethem (2020). *Introduction to machine learning*. MIT press (*p.* 24).

Alqurashi, Khalid A, Khalid S Aljabri and Samia A Bokhari (2011). 'Prevalence of diabetes mellitus in a Saudi community'. In: *Annals of Saudi medicine* 31.1, pages 19–23 (*p.* 112).

Alves, Tiago et al. (2018). 'Dynamic prediction of icu mortality risk using domain adaptation'. In: *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, pages 1328–1336 (*pp.* 3, 9).

American Diabetes Association (2014). 'Diagnosis and classification of diabetes mellitus'. In: *Diabetes care* 37.Supplement 1, S81–S90 (*pp.* 19, 21).

An, Jinwon and Sungzoon Cho (2015). 'Variational autoencoder based anomaly detection using reconstruction probability'. In: *Special Lecture on IE* 2, pages 1–18 (*p.* 31).

Anderson, Ariana E et al. (2016). 'Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study'. In: *Journal of biomedical informatics* 60, pages 162–168 (*pp.* 40, 49, 51).

ANSI-ISO (2005). *ISO/DTR 20514: Health informatics–electronic health record—definition, scope and context* (*pp.* 2, 62).

Atherton, Jim (2011). 'Development of the electronic health record'. In: *AMA Journal of Ethics* 13.3, pages 186–189 (*p.* 62).

Austin, Peter C and Ewout W Steyerberg (2012). 'Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable'. In: *BMC medical research methodology* 12.1, page 82 (*pp.* 36, 97, 100).

Awad, Aya et al. (2017). 'Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach'. In: *International journal of medical informatics* 108, pages 185–195 (*pp.* 3, 9, 21–23, 33, 55, 57, 58).

Baan, Caroline A et al. (1999). 'Performance of a predictive model to identify undiagnosed diabetes in a health care setting.' In: *Diabetes care* 22.2, pages 213–219 (*pp.* 47, 106, 107).

Bassiony, Medhat M et al. (2009). 'Smoking in Saudi Arabia'. In: *Saudi Med J* 30.7, pages 876–81 (*p.* 107).

Batista, Gustavo EAPA, Ronaldo C Prati and Maria Carolina Monard (2004). 'A study of the behavior of several methods for balancing machine learning training data'. In: *ACM SIGKDD explorations newsletter* 6.1, pages 20–29 (*pp.* 47, 64, 106, 112).

Baur, Christoph et al. (2018). 'Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images'. In: *arXiv preprint arXiv:1804.04488* (*p.* 83).

Beagley, Jessica et al. (2014). 'Global estimates of undiagnosed diabetes in adults'. In: *Diabetes research and clinical practice* 103.2, pages 150–160 (*p.* 8).

Benesty, Jacob et al. (2009). 'Pearson correlation coefficient'. In: *Noise reduction in speech processing*. Springer, pages 1–4 (*p.* 85).

Bengio, Yoshua, Ian Goodfellow and Aaron Courville (2017). *Deep learning*. Volume 1. MIT press Massachusetts, USA: (*p.* 27).

Bengio, Yoshua, Patrice Simard and Paolo Frasconi (1994). 'Learning long-term dependencies with gradient descent is difficult'. In: *IEEE transactions on neural networks* 5.2, pages 157–166 (*p.* 29).

Boffetta, Paolo et al. (2011). 'Body mass index and diabetes in Asia: a cross-sectional pooled analysis of 900,000 individuals in the Asia cohort consortium'. In: *PloS one* 6.6, e19930 (*pp.* 107, 136).

Bonora, Enzo and Jaakko Tuomilehto (2011). 'The pros and cons of diagnosing diabetes with A1C'. In: *Diabetes care* 34.Supplement 2, S184–S190 (*pp.* 9, 21).

Boo, YooKyung and YoungJin Choi (2020). 'Comparing logistic regression models with alternative machine learning methods to predict the risk of drug intoxication mortality'. In: *International journal of environmental research and public health* 17.3, page 897 (*p.* 54).

Botsis, Taxiarchis et al. (2010). 'Secondary use of EHR: data quality issues and informatics opportunities'. In: *Summit on Translational Bioinformatics* 2010, page 1 (*p.* 63).

Brajer, Nathan et al. (2020). 'Prospective and External Evaluation of a Machine Learning Model to Predict In-Hospital Mortality of Adults at Time of Admission'. In: *JAMA Network Open* 3.2, e1920733–e1920733 (*pp.* 55, 57–59).

Breiman, Leo (2001). 'Random forests'. In: *Machine learning* 45.1, pages 5–32 (*pp.* 33, 38).

Cao, Xi Hang, Ivan Stojkovic and Zoran Obradovic (2016). 'A robust data scaling algorithm to improve classification accuracies in biomedical data'. In: *BMC bioinformatics* 17.1, pages 1–10 (*p.* 144).

Caruana, Rich, Shumeet Baluja and Tom Mitchell (1996). 'Using the future to" sort out" the present: Rankprop and multitask learning for medical risk evaluation'. In: *Advances in neural information processing systems*, pages 959–965 (*p.* 54).

Celi, Leo Anthony et al. (2012). 'A database-driven decision support system: customized mortality prediction'. In: *Journal of personalized medicine* 2.4, pages 138–148 (*p.* 54).

Centers for Disease Control and Prevention (2020). 'National diabetes statistics report, 2020'. In: *Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services* (*p.* 106).

Central Department of Statistics & Information (CDSI) (2018). *Statistical yearbook 2018*. http://www.cdsi.gov.sa (*p.* 67).

Chalapathy, Raghavendra, Aditya Krishna Menon and Sanjay Chawla (2018). 'Anomaly Detection using One-Class Neural Networks'. In: *arXiv preprint arXiv:1802.06360* (*p.* 31).

Chawla, Nitesh V et al. (2002). 'SMOTE: synthetic minority over-sampling technique'. In: *Journal of artificial intelligence research* 16, pages 321–357 (*pp.* 141, 143, 144).

Cho, Kyunghyun et al. (2014). 'Learning phrase representations using RNN encoder-decoder for statistical machine translation'. In: *arXiv preprint arXiv:1406.1078* (*p.* 30).

Choi, Edward et al. (2016). 'Doctor ai: Predicting clinical events via recurrent neural networks'. In: *Machine Learning for Healthcare Conference*, pages 301–318 (*pp.* 30, 34, 44, 48–51).

Choi, Soo Beom et al. (2014). 'Screening for prediabetes using machine learning models'. In: *Computational and mathematical methods in medicine* 2014 (*pp.* 46, 52, 53).

Christensen, Tom and Anders Grimsmo (2008). 'Instant availability of patient records, but diminished availability of patient information: a multi-method study of GP's use of electronic patient records'. In: *BMC medical informatics and decision making* 8.1, pages 1–8 (*p.* 65).

Chung, Junyoung et al. (2014). 'Empirical evaluation of gated recurrent neural networks on sequence modeling'. In: *arXiv preprint arXiv:1412.3555* (*p.* 30).

Clermont, Gilles et al. (2001). 'Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models'. In: *Critical care medicine* 29.2, pages 291–296 (*p.* 54).

Coorevits, Pascal et al. (2013). 'Electronic health records: new opportunities for clinical research'. In: *Journal of internal medicine* 274.6, pages 547–560 (*pp.* 2, 3, 24, 63).

Cro, Suzie et al. (2020). 'Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: A practical guide'. In: *Statistics in Medicine* (*pp.* 85, 158).

Cunningham, Pádraig, Matthieu Cord and Sarah Jane Delany (2008). 'Supervised learning'. In: *Machine learning techniques for multimedia*. Springer, pages 21–49 (*p.* 26).

Da Silva, Fabio QB et al. (2014). 'Replication of empirical studies in software engineering research: a systematic mapping study'. In: *Empirical Software Engineering* 19.3, pages 501–557 (*p.* 95).

Dai, Andrew M and Quoc V Le (2015). 'Semi-supervised sequence learning'. In: *Advances in Neural Information Processing Systems*, pages 3079–3087 (*p.* 44).

Delahanty, Ryan J, David Kaufman and Spencer S Jones (2018). 'Development and evaluation of an automated machine learning algorithm for in-hospital mortality risk adjustment among critical care patients'. In: *Critical care medicine* 46.6, e481–e488 (*pp.* 55, 57, 58).

Dempster, Arthur P, Nan M Laird and Donald B Rubin (1977). 'Maximum likelihood from incomplete data via the EM algorithm'. In: *Journal of the royal statistical society. Series B (methodological)*, pages 1–38 (*p.* 39).

Deng, Li and Dong Yu (2014). 'Deep learning: methods and applications'. In: *Foundations and trends in signal processing* 7.3–4, pages 197–387 (*p.* 25).

Desjardins, Jeff (2018). *How Big Data Will Unlock the Potential of Healthcare*. https://www.visualcapitalist.com/big-data-healthcare/ (*pp.* 62, 63).

Dey, Rajeeb et al. (2008). 'Application of artificial neural network (ANN) technique for diagnosing diabetes mellitus'. In: *2008 IEEE Region 10 and the Third international Conference on Industrial and Information Systems*. IEEE, pages 1–4 (*pp.* 41, 49, 51).

Doig, GS et al. (1993). 'Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression.' In: *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, page 361 (*p.* 22).

Domingos, Pedro and Michael Pazzani (1997). 'On the optimality of the simple Bayesian classifier under zero-one loss'. In: *Machine learning* 29.2, pages 103–130 (*p.* 38).

DuBose, Katrina D et al. (2012). 'Peer Reviewed: Development and Validation of a Tool for Assessing Glucose Impairment in Adolescents'. In: *Preventing chronic disease* 9 (*p.* 50).

Durga, S, Rishabh Nag and Esther Daniel (2019). 'Survey on machine learning and deep learning algorithms used in internet of things (IoT) healthcare'. In: *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, pages 1018–1022 (*p.* 3).

Dwivedi, Ashok Kumar (2018). 'Analysis of computational intelligence techniques for diabetes mellitus prediction'. In: *Neural Computing and Applications* 30.12, pages 3837–3845 (*pp.* 33, 49, 51).

Edelman, David et al. (2004). 'Utility of hemoglobin A1c in predicting diabetes risk'. In: *Journal of general internal medicine* 19.12, pages 1175–1180 (*pp.* 9, 21).

Elhadd, Tarik A, Abdallah A Al-Amoudi and Ali S Alzahrani (2007). 'Epidemiology, clinical and complications profile of diabetes in Saudi Arabia: a review'. In: *Annals of Saudi medicine* 27.4, pages 241–250 (*pp.* 107, 159).

Esteban, Santiago et al. (2017). 'Development and validation of various phenotyping algorithms for Diabetes Mellitus using data from electronic health records'. In: *Computer methods and programs in biomedicine* 152, pages 53–70 (*pp.* 44, 49, 51).

Evans, RS (2016). 'Electronic health records: then, now, and in the future'. In: *Yearbook of medical informatics* 25.S 01, S48–S61 (*pp.* 2, 62).

Faisal, Muhammad et al. (2020). 'A comparison of logistic regression models with alternative machine learning methods to predict the risk of in-hospital mortality in emergency medical admissions via external validation'. In: *Health informatics journal* 26.1, pages 34–44 (*pp.* 33, 55, 57–59).

Fung, Glenn M and Olvi L Mangasarian (2005). 'Multicategory proximal support vector machine classifiers'. In: *Machine learning* 59.1-2, pages 77–97 (*p.* 42).

Galar, Mikel et al. (2011). 'A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches'. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42.4, pages 463–484 (*pp.* 47, 58, 64).

Gao, Shenghua et al. (2015). 'Single sample face recognition via learning deep supervised autoencoders'. In: *IEEE Transactions on Information Forensics and Security* 10.10, pages 2108–2118 (*p.* 83).

Gao, Yue et al. (2020). 'Machine learning based early warning system enables accurate mortality risk prediction for COVID-19'. In: *Nature communications* 11.1, pages 1–10 (*p.* 59).

Gardner, Matt W and SR Dorling (1998). 'Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences'. In: *Atmospheric Environment* 32.14-15, pages 2627–2636 (*p.* 26).

Gerstein, Hertzel C et al. (2007). 'Annual incidence and relative risk of diabetes in people with various categories of dysglycemia: a systematic overview and meta-analysis of prospective studies'. In: *Diabetes research and clinical practice* 78.3, pages 305–312 (*p.* 20).

Ghassemi, Marzyeh et al. (2015). 'A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data.' In: *AAAI*, pages 446–453 (*p.* 54).

Gilani, Mahryar Taghavi, Majid Razavi and Azadeh Mokhtari Azad (2014). 'A comparison of Simplified Acute Physiology Score II, Acute Physiology and Chronic Health Evaluation II and Acute Physiology and Chronic Health Evaluation III scoring system in predicting mortality and length of stay at surgical intensive care unit'. In: *Nigerian medical journal: journal of the Nigeria Medical Association* 55.2, page 144 (*p.* 23).

Goldberger, Ary L et al. (2000). 'PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals'. In: *circulation* 101.23, e215–e220 (*p.* 55).

Goldenberg, S Larry, Guy Nir and Septimiu E Salcudean (2019). 'A new era: artificial intelligence and machine learning in prostate cancer'. In: *Nature Reviews Urology* 16.7, pages 391–403 (*pp.* 2, 58).

Goldstein, Benjamin A et al. (2017). 'Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review'. In: *Journal of the American Medical Informatics Association* 24.1, pages 198–208 (*p.* 5).

Gómez, Omar S, Natalia Juristo and Sira Vegas (2010). 'Replications types in experimental disciplines'. In: *Proceedings of the 2010 ACM-IEEE international symposium on empirical software engineering and measurement*, pages 1–10 (*p.* 95).

Gondara, L. (Dec. 2016). 'Medical Image Denoising Using Convolutional Denoising Autoencoders'. In: *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 241–246. DOI: 10.1109/ICDMW.2016.0041 (*p.* 83).

Goodacre, Steve, Mike Campbell and Angela Carter (2015). 'What do hospital mortality rates tell us about quality of care?' In: *Emergency Medicine Journal* 32.3, pages 244–247 (*p.* 22).

Goodfellow, Ian et al. (2016). *Deep learning*. Volume 1. MIT press Cambridge (*pp.* 3, 24, 25, 27, 31, 83).

Gray, LJ et al. (2012). 'Detection of impaired glucose regulation and/or type 2 diabetes mellitus, using primary care electronic data, in a multiethnic UK community setting'. In: *Diabetologia* 55.4, pages 959–966 (*pp.* 33, 50, 52, 53).

Greenes, Robert A et al. (2018). 'Clinical decision support models and frameworks: seeking to address research issues underlying implementation successes and failures'. In: *Journal of biomedical informatics* 78, pages 134–143 (*p.* 154).

Griffin, SJ et al. (2000). 'Diabetes risk score: towards earlier detection of type 2 diabetes in general practice'. In: *Diabetes/metabolism research and reviews* 16.3, pages 164–171 (*pp.* 47, 106, 107).

Group, UK Prospective Diabetes Study (1998). 'Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38'. In: *Bmj* 317.7160, pages 703–713 (*p.* 21).

Gu, Qiong et al. (2008). 'Data mining on imbalanced data sets'. In: *2008 International Conference on Advanced Computer Theory and Engineering*. IEEE, pages 1020–1024 (*p.* 47).

Hall, Margaret Jean, Shaleah Levant and Carol J DeFrances (2013). *Trends in inpatient hospital deaths: national hospital discharge survey, 2000-2010*. 118. US Department of Health and Human Services, Centers for Disease Control and ... (*p.* 22).

Hall, Mark and Lloyd Smith (1998). 'Practical feature subset selection for machine learning'. In: (*p.* 153).

Hall, Mark A (2000). 'Correlation-based feature selection of discrete and numeric class machine learning'. In: (*p.* 42).

Han, Hui, Wen-Yuan Wang and Bing-Huan Mao (2005). 'Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning'. In: *International Conference on Intelligent Computing*. Springer, pages 878–887 (*pp.* 141, 144).

Handlos, LN et al. (2013). 'Risk scores for diabetes and impaired glycaemia in the Middle East and North Africa'. In: *Diabetic medicine* 30.4, pages 443–451 (*pp.* 52, 53).

Harerimana, Gaspard et al. (2019). 'Deep learning for electronic health records analytics'. In: *IEEE Access* 7, pages 101245–101259 (*pp.* 13, 63, 156).

Harrell Jr, Frank E, Kerry L Lee and Daniel B Mark (1996). 'Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors'. In: *Statistics in medicine* 15.4, pages 361–387 (*p.* 95).

Harris, Alex HS (2017). 'Path from predictive analytics to improved patient outcomes: a framework to guide use, implementation, and evaluation of accurate surgical predictive models'. In: *Annals of surgery* 265.3, page 461 (*p.* 154).

Hartskamp, Michael van et al. (2019). 'Artificial intelligence in clinical health care applications'. In: *Interactive Journal of Medical Research* 8.2, e12100 (*p.* 2).

Harutyunyan, Hrayr et al. (2019). 'Multitask learning and benchmarking with clinical time series data'. In: *Scientific data* 6.1, pages 1–18 (*pp.* 56–59).

Hasan, Md Kamrul et al. (2020). 'Diabetes prediction using ensembling of different machine learning classifiers'. In: *IEEE Access* 8, pages 76516–76531 (*p.* 2).

Häyrinen, Kristiina, Kaija Saranto and Pirkko Nykänen (2008). 'Definition, structure, content, use and impacts of electronic health records: a review of the research literature'. In: *International journal of medical informatics* 77.5, pages 291–304 (*pp.* 2, 62).

He, Miao et al. (2019). 'CausalBG: Causal Recurrent Neural Network for the Blood Glucose Inference with IoT Platform'. In: *IEEE Internet of Things Journal* 7.1, pages 598–610 (*p.* 59).

Healthcare Analytics Market (2020). https://www.marketsandmarkets.com/Market-Reports/healthcare-data-analytics-market-905.html (*p.* 62).

Hecht, Jeff (2019). 'The future of electronic health records.' In: *Nature* 573.7775, S114 (*p.* 66).

Hill, Tim et al. (1994). 'Artificial neural network models for forecasting and decision making'. In: *International Journal of Forecasting* 10.1, pages 5–15 (*p.* 40).

Hippisley-Cox, Julia et al. (2009). 'Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore'. In: *Bmj* 338, b880 (*pp.* 38, 49, 51).

Hische, Manuela et al. (2010). 'Decision trees as a simple-to-use and reliable tool to identify individuals with impaired glucose metabolism or type 2 diabetes mellitus'. In: *European journal of endocrinology* 163.4, page 565 (*p.* 50).

Hochreiter, Sepp and Jürgen Schmidhuber (1997). 'Long short-term memory'. In: *Neural computation* 9.8, pages 1735–1780 (*pp.* 28, 29, 83, 141).

Holder, Christopher (2018). 'On Semantic Segmentation and Path Planning for Autonomous Vehicles within Off-Road Environments'. PhD thesis. Durham University (*p.* 25).

Hu, Frank B (2011). 'Globalization of diabetes: the role of diet, lifestyle, and genes'. In: *Diabetes care* 34.6, pages 1249–1257 (*p.* 108).

Huang, Guang-Bin et al. (2012). 'Extreme learning machine for regression and multiclass classification'. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.2, pages 513–529 (*pp.* 42, 49, 51).

International Diabetes Federation (2015). *IDF Diabetes Atlas*. http://www.diabetesatlas.org (*pp.* 8, 16, 17).

— (2017). *IDF Diabetes Atlas*. http://www.diabetesatlas.org (*pp.* 16, 17, 21).

— (2019). *IDF Diabetes Atlas*. http://www.diabetesatlas.org (*pp.* 8, 17–19).

International Expert Committee (2009). 'International Expert Committee report on the role of the A1C assay in the diagnosis of diabetes'. In: *Diabetes care* 32.7, pages 1327–1334 (*p.* 21).

Islam, Md Shafiqul, Marwa K Qaraqe and Samir B Belhaouari (2020). 'Early Prediction of Hemoglobin Alc: A novel Framework for better Diabetes Management'. In: *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, pages 542–547 (*pp.* 20, 59).

Jang, Hyun-Jong and Kyung-Ok Cho (2019). 'Applications of deep learning for the analysis of medical data'. In: *Archives of pharmacal research* 42.6, pages 492–504 (*p.* 3).

Jang, Jin-Su, Min-Jun Lee and Tae-Ro Lee (2019). 'Development of T2DM Prediction Model Using RNN'. In: *Journal of Digital Convergence* 17.8, pages 249–255 (*p.* 59).

Japkowicz, Nathalie and Shaju Stephen (2002). 'The class imbalance problem: A systematic study'. In: *Intelligent data analysis* 6.5, pages 429–449 (*p.* 65).

Jayalakshmi, T and A Santhakumaran (2011). 'Statistical Normalization and Back Propagationfor Classification'. In: *International Journal of Computer Theory and Engineering* 3.1, page 89 (*pp.* 42, 49, 51).

Jazayeri, Ali, Ou Stella Liang and Christopher C Yang (2020). 'Imputation of Missing Data in Electronic Health Records Based on Patients' Similarities'. In: *Journal of Healthcare Informatics Research* 4, pages 295–307 (*p.* 159).

Johnson, Alistair EW et al. (2016). 'MIMIC-III, a freely accessible critical care database'. In: *Scientific data* 3, page 160035 (*p.* 56).

Kälsch, Julia et al. (2015). 'Normal liver enzymes are correlated with severity of metabolic syndrome in a large population based cohort'. In: *Scientific reports* 5, page 13058 (*p.* 21).

Karegowda, Asha Gowda, AS Manjunath and MA Jayaram (2011). 'Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes'. In: *International Journal on Soft Computing* 2.2, pages 15–23 (*pp.* 42, 49, 51).

Karpathy, Andrej et al. (2016). 'Cs231n convolutional neural networks for visual recognition'. In: *Neural networks* 1.1 (*p.* 27).

Kaur, Gaganjot and Amit Chhabra (2014). 'Improved J48 classification algorithm for the prediction of diabetes'. In: *International Journal of Computer Applications* 98.22 (*pp.* 39, 40, 49, 51).

Kazemi, Elahe et al. (2014). 'Predicting of trend of hemoglobin a1c in type 2 diabetes: a longitudinal linear mixed model'. In: *International journal of preventive medicine* 5.10, page 1274 (*pp.* 46, 50, 52, 53, 95).

Keegan, Mark T, Ognjen Gajic and Bekele Afessa (2012). 'Comparison of APACHE III, APACHE IV, SAPS 3, and MPM0III and influence of resuscitation status on model performance'. In: *Chest* 142.4, pages 851–858 (*p.* 22).

Keogh, Eamonn et al. (2001). 'Locally adaptive dimensionality reduction for indexing large time series databases'. In: *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, pages 151–162 (*p.* 118).

Khaw, Kay-Tee et al. (2004). 'Association of hemoglobin A1c with cardiovascular disease and mortality in adults: the European prospective investigation into cancer in Norfolk'. In: *Annals of internal medicine* 141.6, pages 413–420 (*pp.* 8, 21).

Kinmonth, Ann-Louise, Nicki Spiegal and Alison Woodcock (1996). 'Developing a training programme in patient-centred consulting for evaluation in a randomised controlled trial; diabetes care from diagnosis in British primary care'. In: *Patient education and counseling* 29.1, pages 75–86 (*p.* 106).

Knaus, William A et al. (1991). 'The APACHE III prognostic system: risk prediction of hospital mortality for critically III hospitalized adults'. In: *Chest* 100.6, pages 1619–1636 (*p.* 22).

Koenig, Ronald J et al. (1976). 'Correlation of glucose regulation and hemoglobin AIc in diabetes mellitus'. In: *New England Journal of Medicine* 295.8, pages 417–420 (*pp.* 8, 19).

Kononenko, Igor (2001). 'Machine learning for medical diagnosis: history, state of the art and perspective'. In: *Artificial Intelligence in medicine* 23.1, pages 89–109 (*p.* 38).

Koopman, Anitra DM et al. (2017). 'The association between social jetlag, the metabolic syndrome, and type 2 diabetes mellitus in the general population: the new Hoorn study'. In: *Journal of biological rhythms* 32.4, pages 359–368 (*p.* 46).

Kopitar, Leon et al. (2020). 'Early detection of type 2 diabetes mellitus using machine learning-based prediction models'. In: *Scientific reports* 10.1, pages 1–12 (*p.* 157).

Kramer, Andrew A, Thomas L Higgins and Jack E Zimmerman (2014). 'Comparison of the Mortality Probability Admission Model III, National Quality Forum, and Acute Physiology and Chronic Health Evaluation IV hospital mortality models: implications for national benchmarking'. In: *Critical care medicine* 42.3, pages 544–553 (*p.* 23).

Kuhn, Max, Kjell Johnson et al. (2013). *Applied predictive modeling*. Volume 26. Springer (*p.* 25).

Larsen, Mogens Lytken, Mogens Hørder and Erik F Mogensen (1990). 'Effect of long-term monitoring of glycosylated hemoglobin levels in insulin-dependent diabetes mellitus'. In: *New England Journal of Medicine* 323.15, pages 1021–1025 (*pp.* 19, 82).

Le, JR Gall et al. (1984). 'A simplified acute physiology score for ICU patients.' In: *Critical care medicine* 12.11, pages 975–977 (*p.* 22).

LeCun, Yann, Yoshua Bengio and Geoffrey Hinton (2015). 'Deep learning'. In: *Nature* 521.7553, pages 436–444 (*pp.* 26–28, 31, 83, 110, 141).

Lemeshow, Stanley et al. (1993). 'Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients'. In: *Jama* 270.20, pages 2478–2486 (*p.* 22).

Lichman, M. (2013). *UCI Machine Learning Repository*. URL: http://archive.ics.uci.edu/ml (*pp.* 38, 48).

Lin, Danyu Y and Lee-Jen Wei (1989). 'The robust inference for the Cox proportional hazards model'. In: *Journal of the American statistical Association* 84.408, pages 1074–1078 (*p.* 38).

Lindsay, R Murray and Andrew SC Ehrenberg (1993). 'The design of replicated studies'. In: *The American Statistician* 47.3, pages 217–228 (*pp.* 95, 96).

Lipton, Zachary C (2018). 'The mythos of model interpretability'. In: *Queue* 16.3, pages 31–57 (*p.* 111).

Lipton, Zachary C et al. (2015). 'Learning to diagnose with LSTM recurrent neural networks'. In: *arXiv preprint arXiv:1511.03677* (*pp.* 34, 44, 48, 49, 51).

Liu, Zihao et al. (2019). 'Machine vision guided 3d medical image compression for efficient transmission and accurate segmentation in the clouds'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12687–12696 (*p.* 62).

Longadge, Rushi and Snehalata Dongre (2013). 'Class imbalance problem in data mining review'. In: *arXiv preprint arXiv:1305.1707* (*pp.* 64, 112).

Lundberg, Scott M and Su-In Lee (2017). 'A unified approach to interpreting model predictions'. In: *Advances in neural information processing systems*, pages 4765–4774 (*pp.* 111, 112).

Luo, Yuan et al. (2016). 'Predicting ICU Mortality Risk by Grouping Temporal Trends from a Multivariate Panel of Physiologic Measurements.' In: *AAAI*, pages 42–50 (*pp.* 3, 9, 54, 139).

Maaten, Laurens van der and Geoffrey Hinton (2008). 'Visualizing data using t-SNE'. In: *Journal of machine learning research* 9.Nov, pages 2579–2605 (*pp.* 86, 127, 149).

Manca, Donna P (2015). 'Do electronic medical records improve quality of care?: Yes'. In: *Canadian Family Physician* 61.10, page 846 (*p.* 62).

Mandrekar, Jayawant N (2010). 'Receiver operating characteristic curve in diagnostic test assessment'. In: *Journal of Thoracic Oncology* 5.9, pages 1315–1316 (*p.* 34).

Mani, Subramani et al. (2012). 'Type 2 diabetes risk forecasting from EMR data using machine learning'. In: *AMIA annual symposium proceedings*. Volume 2012. American Medical Informatics Association, page 606 (*pp.* 39, 49, 51).

Mazurowski, Maciej A et al. (2008). 'Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance'. In: *Neural networks* 21.2, pages 427–436 (*pp.* 43, 65).

McCarter, Robert J, James M Hempe and Stuart A Chalew (2006). 'Mean blood glucose and biological variation have greater influence on HbA1c levels than glucose instability: an analysis of data from the Diabetes Control and Complications Trial'. In: *Diabetes Care* 29.2, pages 352–355 (*pp.* 45, 94).

McDonald, John H (2009). *Handbook of biological statistics*. Volume 2. sparky house publishing Baltimore, MD (*p.* 120).

Meng, Xue-Hui et al. (2013). 'Comparison of three data mining models for predicting diabetes or prediabetes by risk factors'. In: *The Kaohsiung journal of medical sciences* 29.2, pages 93–99 (*pp.* 41, 43, 49, 51).

Meystre, Stephane M et al. (2017). 'Clinical data reuse or secondary use: current status and potential future progress'. In: *Yearbook of medical informatics* 26.1, page 38 (*p.* 63).

Miotto, Riccardo et al. (2016). 'Deep patient: An unsupervised representation to predict the future of patients from the electronic health records'. In: *Scientific reports* 6, page 26094 (*pp.* 2, 34, 44, 49, 51, 56, 133).

Moon, Sanghee et al. (2020). 'Classification of Parkinson's disease and essential tremor based on balance and gait characteristics from wearable motion sensors via machine learning techniques: a data-driven approach'. In: *Journal of NeuroEngineering and Rehabilitation* 17.1, pages 1–8 (*p.* 34).

Motka, Rakesh et al. (2013). 'Diabetes mellitus forecast using different data mining techniques'. In: *Computer and Communication Technology (ICCCT), 2013 4th International Conference on*. IEEE, pages 99–103 (*pp.* 43, 49, 51).

Naqvi, Syeda et al. (2017). 'Correlation between glycated hemoglobin and triglyceride level in type 2 diabetes mellitus'. In: *Cureus* 9.6 (*p.* 114).

Nashef, Samer AM et al. (1999). 'European system for cardiac operative risk evaluation (Euro SCORE)'. In: *European journal of cardio-thoracic surgery* 16.1, pages 9–13 (*p.* 23).

Nathan, David M et al. (2008). 'Translating the A1C assay into estimated average glucose values'. In: *Diabetes care* 31.8, pages 1473–1478 (*pp.* 45, 94).

Nembrini, Stefano, Inke R König and Marvin N Wright (2018). 'The revival of the Gini importance?' In: *Bioinformatics* 34.21, pages 3711–3718 (*p.* 135).

Ngufor, Che et al. (2019). 'Mixed Effect Machine Learning: a framework for predicting longitudinal change in hemoglobin A1c'. In: *Journal of biomedical informatics* 89, pages 56–67 (*p.* 46).

Nguyen, Hien M, Eric W Cooper and Katsuari Kamei (2011). 'Borderline over-sampling for imbalanced data classification'. In: *International Journal of Knowledge Engineering and Soft Data Paradigms* 3.1, pages 4–21 (*p.* 144).

Noble, William S (2006). 'What is a support vector machine?' In: *Nature Biotechnology* 24.12, pages 1565–1567 (*p.* 33).

Nosek, Brian A and Timothy M Errington (2020). 'What is replication?' In: *PLoS biology* 18.3, e3000691 (*p.* 95).

Panesar, Arjun (2019). *Machine learning and AI for healthcare*. Springer (*p.* 156).

Pangaribuan, Jefri Junifer et al. (2014). 'Diagnosis of diabetes mellitus using extreme learning machine'. In: *2014 International Conference on Information Technology Systems and Innovation (ICITSI)*. IEEE, pages 33–38 (*p.* 42).

Payrovnaziri, Seyedeh Neelufar et al. (2020). 'The Impact of Missing Value Imputation on the Interpretations of Predictive Models: A Case Study on One-year Mortality Prediction in ICU Patients with Acute Myocardial Infarction'. In: *medRxiv* (*p.* 159).

Perveen, Sajida et al. (2019). 'Prognostic modeling and prevention of diabetes using machine learning technique'. In: *Scientific reports* 9.1, pages 1–9 (*pp.* 59, 159).

Peterson, Karen P et al. (1998). 'What is hemoglobin A1c? An analysis of glycated hemoglobins by electrospray ionization mass spectrometry'. In: *Clinical Chemistry* 44.9, pages 1951–1958 (*pp.* 8, 19).

Pivovarov, Rimma (2015). 'Electronic health record summarization over heterogeneous and irregularly sampled clinical data'. PhD thesis. Columbia University (*pp.* 65, 66).

Polat, Kemal and Salih Güneş (2007). 'An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease'. In: *Digital Signal Processing* 17.4, pages 702–710 (*p.* 43).

Polat, Kemal, Salih Güneş and Ahmet Arslan (2008). 'A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine'. In: *Expert systems with applications* 34.1, pages 482–487 (*pp.* 38, 41, 49, 51).

Pradhan, Aruna D et al. (2007). 'Hemoglobin A1c predicts diabetes but not cardiovascular disease in nondiabetic women'. In: *The American journal of medicine* 120.8, pages 720–727 (*pp.* 19, 21, 82).

Pradhan, Manaswini and Ranjit Kumar Sahu (2011). 'Predict the onset of diabetes disease using Artificial Neural Network (ANN)'. In: *International Journal of Computer Science & Emerging Technologies (E-ISSN: 2044-6004)* 2.2 (*pp.* 42, 49, 51).

Priyadarshini, Rojalina, Nilamadhab Dash and Rachita Mishra (2014). 'A Novel approach to predict diabetes mellitus using modified Extreme learning machine'. In: *Electronics and Communication Systems (ICECS), 2014 International Conference on.* IEEE, pages 1–5 (*pp.* 42, 49, 51).

Provost, Foster (2000). 'Machine learning from imbalanced data sets 101'. In: *Proceedings of the AAAI'2000 workshop on imbalanced data sets.* Volume 68. 2000. AAAI Press, pages 1–3 (*p.* 47).

Rahman, M Mostafizur and Darryl N Davis (2013). 'Addressing the class imbalance problem in medical datasets'. In: *International Journal of Machine Learning and Computing* 3.2, page 224 (*pp.* 47, 64, 65, 112).

Rahmanian, Karamatollah et al. (2016). 'The association between pre-diabetes with body mass index and marital status in an Iranian urban population'. In: *Global journal of health science* 8.4, page 95 (*p.* 136).

Rajkomar, Alvin et al. (2018). 'Scalable and accurate deep learning with electronic health records'. In: *NPJ Digital Medicine* 1.1, pages 1–10 (*p.* 135).

Rallapalli, Sreekanth and T Suryakanthi (2016). 'Predicting the risk of diabetes in big data electronic health Records by using scalable random forest classification algorithm'. In: *2016 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. IEEE, pages 281–284 (*pp.* 40, 49, 51).

Rau, Hsiao-Hsien et al. (2016). 'Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network'. In: *Computer methods and programs in biomedicine* 125, pages 58–65 (*pp.* 44, 49, 51).

Raudys, Sarunas J, Anil K Jain et al. (1991). 'Small sample size effects in statistical pattern recognition: Recommendations for practitioners'. In: *IEEE Transactions on pattern analysis and machine intelligence* 13.3, pages 252–264 (*p.* 43).

Rawlings, John O, Sastry G Pantula and David A Dickey (2001). *Applied regression analysis: a research tool*. Springer Science & Business Media (*p.* 33).

Razzak, Muhammad Imran, Muhammad Imran and Guandong Xu (2020). 'Big data analytics for preventive medicine'. In: *Neural Computing and Applications* 32.9, pages 4417–4451 (*p.* 6).

Ribeiro, Marco Tulio, Sameer Singh and Carlos Guestrin (2016). '" Why should I trust you?" Explaining the predictions of any classifier'. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144 (*pp.* 111, 112).

Robinson, CA, G Agarwal and K Nerenberg (2011). 'Validating the CANRISK prognostic model for assessing diabetes risk in Canada's multi-ethnic population'. In: *Chronic diseases and injuries in Canada* 32.1 (*p.* 50).

Rose, Eric and Debra S Ketchell (2003). 'Does daily monitoring of blood glucose predict hemoglobin A1c levels?' In: *Clinical Inquiries, 2003 (MU)* (*pp.* 45, 95).

Rumelhart, David E and James L McClelland (1986). 'Parallel distributed processing: explorations in the microstructure of cognition. volume 1. foundations'. In: (*p.* 31).

Saeed, Mohammed et al. (2011). 'Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database'. In: *Critical care medicine* 39.5, page 952 (*p.* 54).

Saito, Takaya and Marc Rehmsmeier (2015). 'The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets'. In: *PloS one* 10.3, e0118432 (*p.* 101).

Salman, Rabha AbdulAziz, Adel Salman AlSayyad and Craig Ludwig (2019). 'Type 2 diabetes and healthcare resource utilisation in the Kingdom of Bahrain'. In: *BMC health services research* 19.1, page 939 (*p.* 18).

Sanghani, Nandita B et al. (2013). 'Impact of lifestyle modification on glycemic control in patients with type 2 diabetes mellitus'. In: *Indian journal of endocrinology and metabolism* 17.6, page 1030 (*p.* 85).

Sarwar, Abid and Vinod Sharma (2014). 'Comparative analysis of machine learning techniques in prognosis of type II diabetes'. In: *AI & society* 29.1, pages 123–129 (*pp.* 43, 49, 51).

Schmidt, Jonathan et al. (2019). 'Recent advances and applications of machine learning in solid-state materials science'. In: *npj Computational Materials* 5.1, pages 1–36 (*p.* 25).

Schuster, Mike and Kuldip K Paliwal (1997). 'Bidirectional recurrent neural networks'. In: *IEEE transactions on Signal Processing* 45.11, pages 2673–2681 (*p.* 30).

Schütze, Hinrich, Christopher D Manning and Prabhakar Raghavan (2008). *Introduction to information retrieval*. Volume 39. Cambridge University Press Cambridge (*pp.* 34, 35).

Al-Shayea, Qeethara Kadhim (2011). 'Artificial neural networks in medical diagnosis'. In: *International Journal of Computer Science Issues* 8.2, pages 150–154 (*p.* 40).

Shi, Hon-Yi et al. (2012). 'Comparison of artificial neural network and logistic regression models for predicting in-hospital mortality after primary liver cancer surgery'. In: *PloS one* 7.4, e35781 (*p.* 54).

Shin, H. et al. (Dec. 2011). 'Autoencoder in Time-Series Analysis for Unsupervised Tissues Characterisation in a Large Unlabelled Medical Image Dataset'. In: *2011 10th International Conference on Machine Learning and Applications and Workshops*. Volume 1, pages 259–264. DOI: 10.1109/ICMLA.2011.38 (*p.* 83).

Silva, Thiago JA et al. (2009). 'Predictors of in-hospital mortality among older patients'. In: *Clinics* 64.7, pages 613–618 (*p.* 54).

Sperandei, Sandro (2014). 'Understanding logistic regression analysis'. In: *Biochemia Medica* 24.1, pages 12–18 (*p.* 33).

Steck, Harald (2020). 'Autoencoders that don't overfit towards the identity'. In: *34th Conference on Neural Information Processing Systems (NeurIPS)* (*p.* 32).

Stratton, Irene M et al. (2000). 'Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study'. In: *Bmj* 321.7258, pages 405–412 (*p.* 21).

Sun, Chen et al. (2017). 'Revisiting unreasonable effectiveness of data in deep learning era'. In: *Proceedings of the IEEE international conference on computer vision*, pages 843–852 (*p.* 47).

Suykens, Johan AK and Joos Vandewalle (1999). 'Least squares support vector machine classifiers'. In: *Neural processing letters* 9.3, pages 293–300 (*pp.* 38, 42).

Szandała, Tomasz (2021). 'Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks'. In: *Bio-inspired Neurocomputing*. Springer, pages 203–224 (*p.* 27).

Tang, Jiliang, Salem Alelyani and Huan Liu (2014). 'Feature selection for classification: A review'. In: *Data classification: Algorithms and applications*, page 37 (*p.* 55).

Tang, Yuchun et al. (2009). 'SVMs modeling for highly imbalanced classification'. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.1, pages 281–288 (*p.* 144).

Temurtas, Hasan, Nejat Yumusak and Feyzullah Temurtas (2009). 'A comparative study on diabetes disease diagnosis using neural networks'. In: *Expert Systems with applications* 36.4, pages 8610–8615 (*pp.* 41, 43, 49, 51).

Urdan, Timothy C (2016). *Statistics in plain English*. Taylor & Francis (*p.* 127).

Van Der Maaten, Laurens, Eric Postma and Jaap Van den Herik (2009). 'Dimensionality Reduction: A Comparative Review'. In: *J Mach Learn Res* 10, pages 66–71 (*p.* 31).

Vapnik, Vladimir (2013). *The nature of statistical learning theory*. Springer Science & Business Media (*p.* 33).

Velu, CM and KR Kashwan (2013). 'Visual data mining techniques for classification of diabetic patients'. In: *Advance Computing Conference (IACC), 2013 IEEE 3rd International*. IEEE, pages 1070–1075 (*pp.* 39, 49, 51).

Venkatesan, P and S Anitha (2006). 'Application of a radial basis function neural network for diagnosis of diabetes mellitus'. In: *Current Science* 91.9, pages 1195–1199 (*pp.* 40, 49, 51).

Vijayan, V Veena and C Anjali (2015). 'Prediction and diagnosis of diabetes mellitus—A machine learning approach'. In: *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. IEEE, pages 122–127 (*p.* 49).

Vincent, Jean-Louis, Arnaldo De Mendonça et al. (1998). 'Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study'. In: *Critical care medicine* 26.11, pages 1793–1800 (*p.* 22).

Vincent, Jean-Louis and Mervyn Singer (2010). 'Critical care: advances and future perspectives'. In: *The Lancet* 376.9749, pages 1354–1361 (*p.* 23).

Vincent, Pascal et al. (2010). 'Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion'. In: *Journal of Machine Learning Research* 11.Dec, pages 3371–3408 (*pp.* 83, 86, 146).

Vogenberg, F Randy (2009). 'Predictive and prognostic models: implications for healthcare decision-making in a modern recession'. In: *American health & drug benefits* 2.6, page 218 (*pp.* 6, 154).

Walker, Richard M, Oliver James and Gene A Brewer (2017). *Replication, experiments and knowledge in public management research* (*p.* 95).

Wang, Ying et al. (2020). 'Prediction of Diabetes Based on Convolutional Neural Network'. In: (*p.* 59).

Weiss, Gary M and Foster Provost (2001). 'The effect of class distribution on classifier learning: an empirical study'. In: *Rutgers Univ* (*p.* 64).

Wells, Brian J, Kevin M Chagin et al. (2013). 'Strategies for handling missing data in electronic health record derived data'. In: *Egems* 1.3 (*p.* 65).

Wells, Brian J, Kristin M Lenoir et al. (2018). 'Predicting current glycated hemoglobin values in adults: development of an algorithm from the electronic health record'. In: *JMIR medical informatics* 6.4, e10780 (*pp.* 2, 34, 46, 52, 53, 77, 95, 107, 110, 114, 155, 158).

Williams, DRR et al. (1995). 'Undiagnosed glucose intolerance in the community: the Isle of Ely Diabetes Project'. In: *Diabetic medicine* 12.1, pages 30–35 (*p.* 106).

Williams, Rhys et al. (2020). 'Global and regional estimates and projections of diabetes-related health expenditure: Results from the International Diabetes Federation Diabetes Atlas'. In: *Diabetes Research and Clinical Practice*, page 108072 (*pp.* 17, 18).

Wise, Jacqui (2014). 'A third of adults in England have "prediabetes," study says'. In: *Bmj* 348 (*p.* 17).

World Health Organisation (2004). *ICD-10 : international statistical classification of diseases and related health problems : tenth revision*. https://apps.who.int/iris/handle/10665/42980, Last accessed on 2020-12-15 (*p.* 68).

— (2011). *Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus: abbreviated report of a WHO consultation*. Technical report. World Health Organisation (*pp.* 8, 20).

— (2014). *Global Status Report on Noncommunicable Diseases*. World Health Organisation (*p.* 17).

— (2016). *Global Report on Diabetes*. http://www.who.int/diabetes/global-report/en/ (*pp.* 2, 17).

— (2020). *ICD-10 : international statistical classification of diseases and related health problems : 11th revision*. https://www.who.int/classifications/classification-of-diseases, Last accessed on 2020-12-15 (*p.* 66).

Wright, John et al. (2006). 'Learning from death: a hospital mortality reduction programme'. In: *Journal of the Royal Society of Medicine* 99.6, pages 303–308 (*p.* 22).

Wu, Jiang et al. (2009). 'A semi-supervised learning based method: Laplacian support vector machine used in diabetes disease diagnosis'. In: *Interdisciplinary Sciences: Computational Life Sciences* 1.2, pages 151–155 (*p.* 38).

Wu, Yanling et al. (2014). 'Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention'. In: *International journal of medical sciences* 11.11, page 1185 (*p.* 16).

Xin, Zhong et al. (2010). 'A simple tool detected diabetes and prediabetes in rural Chinese'. In: *Journal of clinical epidemiology* 63.9, pages 1030–1035 (*p.* 50).

Xu, Stanley et al. (2014). 'Accuracy of hemoglobin A1c imputation using fasting plasma glucose in diabetes research using electronic health records data'. In: *Statistics, Optimization & Information Computing* 2.2, pages 93–104 (*pp.* 33, 45, 50, 52, 53, 56, 132).

Yoon, Kun-Ho et al. (2006). 'Epidemic obesity and type 2 diabetes in Asia'. In: *The Lancet* 368.9548, pages 1681–1688 (*pp.* 107, 108).

Yue-Hei Ng, Joe et al. (2015). 'Beyond short snippets: Deep networks for video classification'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702 (*p.* 44).

Al-Zahrani, Jamaan M et al. (2019). 'Prevalence of Prediabetes, Diabetes and Its Predictors among Females in Alkharj, Saudi Arabia: A Cross-Sectional Study'. In: *Annals of Global Health* 85.1 (*pp.* 133, 158).

Zhang, Liyuan, Huamin Yang and Zhengang Jiang (2018). 'Imbalanced biomedical data classification using self-adaptive multilayer ELM combined with dynamic GAN'. In: *Biomedical engineering online* 17.1, page 181 (*pp.* 64, 112).

Zhang, Xuanping et al. (2010). 'A1C level and future risk of diabetes: a systematic review'. In: *Diabetes care* 33.7, pages 1665–1673 (*p.* 9).

Zhao, Jing et al. (2017). 'Learning from heterogeneous temporal data in electronic health records'. In: *Journal of biomedical informatics* 65, pages 105–119 (*pp.* 62, 118).

Zou, Quan et al. (2018). 'Predicting diabetes mellitus with machine learning techniques'. In: *Frontiers in genetics* 9, page 515 (*pp.* 45, 49, 51).